Essays in Political Methodology

Gary King Matthew Lee Blackwell

Essays in Political Methodology

Abstract

This dissertation provides three novel methodologies to the field of political science. In the first chapter, I describe how to make causal inferences in the face of dynamic strategies. Traditional causal inference methods assume that these dynamic decisions are made all at once, an assumption that forces a choice between omitted variable bias and post-treatment bias. I resolve this dilemma by adapting methods from biostatistics and use these methods to estimate the effectiveness of an inherently dynamic process: a candidate's decision to "go negative." Drawing on U.S. statewide elections (2000-2006), I find, in contrast to the previous literature, that negative advertising is an effective strategy for non-incumbents.

In the second chapter, I develop a method for handling measurement error. Social scientists devote considerable effort to mitigating measurement error during data collection but then ignore the issue during data analysis. Although many statistical methods have been proposed for reducing measurement error-induced biases, few have been widely used because of implausible assumptions, high levels of model dependence, difficult computation, or inapplicability with multiple mismeasured variables. This chapter develops an easy-to-use alternative without these problems; it generalizes the popular multiple imputation framework by treating missing data problems as a special case of extreme measurement error and corrects for both.

In the final chapter, I introduce a model for detecting changepoints in the distribution of contributions to candidates over the course of a campaign. This *game-changers* model is ideal for campaign contributions data because it allows for overdispersion, a key feature of contributions data. While many extant changepoint models force researchers to choose the number of changepoint *ex ante*, the game-changers model incorporates a Dirichlet process prior in order to estimate the number of changepoints along with their location. I demonstrate the usefulness of the model in data from the 2012 Republican primary and the 2008 U.S. Senate elections.

Contents

Author List

The following authors contributed to Chapter 2: James Honaker and Gary King.

List of Figures

To all those who put up with me.

Acknowledgments

As with all work, some thanks are in order: Steve Ansolabehere, Amy Catalinac, Andrew Coe, Adam Glynn, Justin Grimmer, Jens Hainmueller, James Honaker, Luke Keele, Gary King, Burt Monroe, Clayton Nall, Michael Peress, James Robins, Brian Schaffner, Maya Sen, James Snyder, Arthur Spirling, Elizabeth Stuart, Brandon Van Dyck, Jonathan Wand, Teppei Yamamoto, and Chris Zorn. These are people I am happy and lucky to call friends and colleagues. In addition, I received wonderful comments from seminar participants at Stanford University and the University of California, Berkeley. The Institute for Quantitative Social Science at Harvard University provided for many aspects of the work. Any remaining bad jokes (or errors) remain mine, and mine alone.

The only reason for time is so that everything doesn't happen at once.

Albert Einstein

1

A Framework for Dynamic Causal Inference in Political Science

What candidate would plan all of their rallies, write all of their speeches, and film all of their advertisements at beginning of a campaign, then sit back and watch them unfold until election day? Clearly this is absurd, and yet it is the only setup that the usual ways of making causal inferences allows us to study. While political science has seen enormous growth in attention to causal inference over the past decade, these advances have heavily focused on snapshots where the dynamic nature of politics are crammed into a single point in time. As political science finds itself with a growing

number of motion pictures—panel data, time-series cross-sectional data—a tension has emerged between substance and method. Indeed, applied to dynamic data, the best practices of *single-shot* causal inference methods provide conflicting advice and fail to alleviate omitted variable or post-treatment bias.

This essay focuses on a specific dynamic process: negative advertising in 176 U.S. Senate and Gubernatorial elections from 2000 until 2006. Candidates in these races change their tone over the course of the campaign, reacting to their current environment. A single-shot causal inference method compares campaigns that are similar on a host of pre-election variables in order to eliminate omitted variable bias. While this is often the best approach with single-shot data, such an approach ignores the fundamentally dynamic nature of campaigns: races that become close over the course of the campaign are more likely to go negative than those that are safe. Attempting to correct for this dynamic selection by controlling for polls leads to post-treatment bias since earlier campaign tone influences polling. The inappropriate application of single-shot causal inference therefore leaves scholars between a rock and hard place, steeped in bias with either approach. This dilemma is not limited to negative advertising or campaigns—every field of political science has a variable of interest that evolves over time.

This essay solves this dilemma by presenting a framework for dynamic causal inference and a set of tools, developed in biostatistics and epidemiology (Robins, Hernán, and Brumback, 2000), to estimate dynamic causal effects. These tools directly model dynamic selection and overcome the above problems of single-shot causal inference. Actions (such as campaign tone) are allowed to vary over time along with any confounding covariates (such as polling). Thus, we can study the effects of the *action history* (candidate's tone across the entire campaign) as opposed to a single action (simply "going negative").

To estimate dynamic causal effects, this essay applies inverse probability of treatment weighting (IPTW) to class of semi-parametric models called marginal structural models (MSM). These models dictate the form of the relationship between the large number of possible action histories and the outcome and serve to reduce the number of causal parameters. The dynamic causal effects are encoded

as parameters of the MSM and, under certain assumptions, IPTW estimates them free of the biases inherent in single-shot methods. With this approach, each unit is weighted by the inverse of the estimated probability of its observed action history. This weighting creates a pseudosample where dynamic selection is eliminated, circumventing the dilemmas posed by single-shot causal inference. Since these methods require strong assumptions, this essay also develops a novel diagnostic tool, the history-adjusted balance, and describes a sensitivity analysis framework to address potential causes of concern.

Once I correct for the biases due to time, I find that negative advertising is an effective strategy for Democratic non-incumbents. This stands in contrast to the previous literature on negative advertising, which, according to Lau, Sigelman, and Rovner (2007), "does not bear out the idea that negative advertising is an effective means of winning votes." The previous approaches to estimating the effectiveness of negative advertising relied on single-shot methods and when I apply these methods to the present data, I find similar non-effects. These single-shot results are worrisome since, as I show below, polls in the middle of campaign are one of the most important predictors of the decision to go negative. This crucial selection issue is inherently dynamic and has been largely ignored by previous work on this topic.

The essay proceeds as follows. Section 1.1 describes the how dynamic causal inference extends single-shot methods. Section 1.2 introduces marginal structural models and the assumptions that they use. Section 1.3 describes a weighting approach to estimating dynamic causal effects. Section 1.4 applies the techniques to estimating the effectiveness of "going negative" in campaigns. Section 1.5 discusses useful diagnostics and a sensitivity analysis framework for marginal structural models. Section 1.6 concludes with directions for future research.

1.1 Single-shot and dynamic causal inference

The goal of a single-shot approach causal inference is to estimate the effect of a single *action* on an outcome at a single point in time.[1] With the example of campaigns, we might be interested in the effect of a Democratic candidate running a negative campaign or a positive one on his or her share of the two-party vote. There are many situations in political science, including campaigns, where actions evolve over time and react to the current state of affairs. In this case, a campaign can "go negative" at multiple points over the course of the campaign. Perhaps a candidate attacks early, before their opponent has a footing, or perhaps she runs negative ads late, responding to smear tactics. These two situations, as far apart as they are, would both register as "going negative" in a single-shot model since they ignore time and implicitly assume that all actions occur at once. This is an acceptable framework for many problems because actions really do occur once. When actions unfold over time, however, the incorporation of time and its implications become necessary.

1.1.1 Actions versus action sequences

Dynamic causal inference, in contrast, allows the actions to vary over time. In this framework, we investigate the effect of an *action sequence* on the outcome of interest. In this framework, we have sequences such as $(\text{positive}_1, \text{negative}_2)$, where a candidate stays positive in the first part of the campaign and then goes negative later. We might ask how this differs from initiating negativity earlier in the race: $(\text{negative}_1, \text{negative}_2)$. This framework has two primary advantages over single-shot methods in dynamic situations. First, the comparison of action sequences naturally handles a richer set of causal questions, which include both the presence and timing of actions. As Pierson (2000) points out, *when* an action occurs is often as important as *if* it occurs at all. Second, and more important, this framework clarifies and resolves certain dilemmas posed by single-shot methods.

1. Actions here are synonymous with *treatments*, a more common term in the causal inference literature.

1.1.2 Confounders in a single-shot world

In order to separate a causal effect from a mere association, we must be confident that the observed correlations are not due to some other variable. In political science, we call this assumption *no omitted variables* and it is made, implicitly or explicitly, in almost all empirical research in political science. It states that we have measured and appropriately controlled for any variable that could potentially cause bias in our causal estimates. We call this bias *confounding* and the variables that cause it *confounders*. For instance, if we were to run a regression of the Democratic vote share in Senate elections on a measure of Democratic negativity, we would also want to control for variables that might cause a correlation between negativity and vote shares. In this case, the incumbency status of the Democrat would a confounder, because incumbents are less likely to go negative and also more likely to win an election. Figure 1.1a shows a graphical of the causal relationship between confounders, actions, and the outcome. We would want to include confounders like these in our analysis, be it using a linear model, a generalized linear model, or a matching estimator.

How do we choose which of our variables are confounders? A common definition is this: a confounder is any variable that (a) is correlated with the outcome, (b) causes, or shares a common cause with, the action, and (c) is not affected by the action. Thus, in any regression or matching estimation of a causal effect, we would want to control for or match on any *pre-action variables* in the sense that they are causally prior to the action of interest. In negative advertising, these variables would either affect or be correlated with the decision to go negative, but never be affected by the decision to go negative. We avoid controlling for *post-action* variables because doing so can induce bias in estimating the causal effect. This is known in the causal inference literature as *post-treatment bias* (Ho et al., 2006).

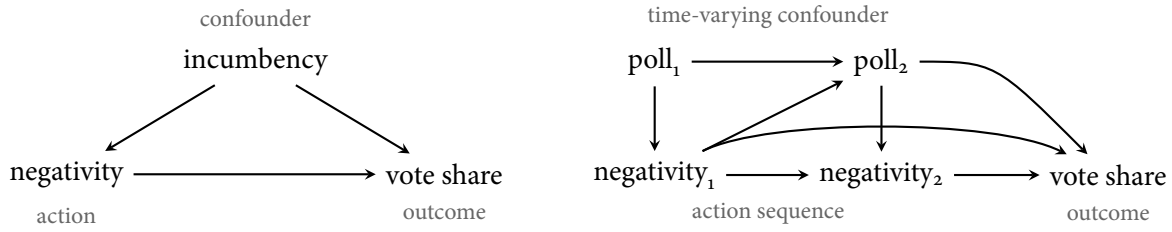
There are two related sources of post-treatment bias. First, conditioning on post-action variables can "block" part of the action's overall effect. For instance, suppose a researcher controlled for polling results from the day of the election when attempting to estimate the effect of incumbency. This will understate the effect of incumbency since most of the effect flows through the standing of candidates

late in the race. Second, conditioning on a post-action variable can induce selection bias even when no bias exist absent the conditioning. For instance, suppose at the start of campaign we randomly assigned high and low budgets to different Democratic candidates for Senate. If we condition on the polls sometime during the campaign, we can seriously bias our estimates of the effect of campaign budgets. Those leading Democrats who had high budgets are likely to differ strongly from leaders with small budgets. For example, if higher budgets help a candidate, then those low-budget leaders are actually much stronger candidates than the high-budget leaders, since they were able to lead in the polls without the additional funding. Thus, comparing high- and low-budget leaders would give a misleading estimate of the causal effect of campaign finance, even though it was randomly assigned.

1.1.3 The problem of time-varying confounders

When we force an inherently dynamic situation into a single-shot framework, the above discussion of confounders and post-treatment bias becomes muddled. Take negative advertising, for example: how should we treat polling data from over the course of the campaign? It is surely *pre-action* in the sense that polling affects the decision to go negative and it is correlated with the outcome, the election outcome. At the same time, polling is *post-action* since it is affected by negativity earlier in the race. Polling is an example of a *time-varying confounder*, which is a confounder that both affects future treatment and is affected by past treatment.

The single-shot advice to include pre-action confounders and exclude post-action variables appears to recommend both courses of action in this situation, leaving a researcher without a palatable solution. In fact, both of these approaches will bias causal estimates, albeit in different ways. In the above hypothetical regression of Democratic vote share on Democratic negativity, we could omit polling data from the regression on the grounds that it is post-treatment, yet this would lead to omitted variable bias. Note that in Figure 1.1b that polling in period 2 affects negativity in period 2, perhaps because candidates that are trailing are more likely to go negative. If we exclude polling from our regression (or



(a) Single-shot causal inference. **(b)** Dynamic causal inference.

Figure 1.1: Directed Acyclic Graphs showing single-shot and dynamic causal inference frameworks. Each arrow represents a causal relationship.

matching analysis), it might seem that negativity is a bad strategy even though this is wholly due to candidates going negative when they are in trouble. Thus, we must include polling in our analyses. Doing so, however, also biases our estimates, since the polls in period 2 are partially a result of negativity in period 1. For instance, a candidate who stays positive early and whose polls decline might have done better if she had gone negative early. If we control for polling in period 2, we block that part of the effect in our analysis and introduce post-treatment bias into our estimates. Either approach, ignoring or including polling, will lead to some form of bias. These problems with time-varying confounders cannot be solved by single-shot methods even if we assume, as I do below, no omitted variables in each time period. They are fundamental to situations where actions unfold over time. The assumptions and methods presented in this essay represent a solution to this dilemma under the weakest possible assumptions on the causal structure of the data. [2]

1.1.4 A weighting approach to dynamic causal inference

A key characteristic of single-shot methods, such as regression and matching, is that they provide no way of removing omitted variable bias due to time-varying confounders without inducing

2. We could change our quantity of interest to avoid this problem, but we would be restricting our analysis to the effect of the last action. Furthermore, the assumptions used to identify this quantity of interest would likely be as strong as the assumptions used to identify the full set of causal quantities.

post-treatment bias. These estimators divide up the data into comparable subsets and estimate causal effects within these subsets. The overall effect is simply a combination of these *stratum-specific* effects. This broad approach is called *stratification* and it breaks down in dynamic settings as described above: stratifying removes omitted variable bias but induces post-treatment bias for time-varying confounders.

An alternative to stratification estimators are *inverse probability of treatment weighting* (IPTW) estimators, which reweight the data to alleviate the omitted variable bias.[3] To see how these weights work, note that, in negative campaigning, certain strategies are used more often than others: candidates tend to go negative when they are trailing and stay positive when they are leading. In Figure 1.1b, we represent this in the arrows from polling to negativity. This, of course, causes confounding. To remove this time-varying confounding, we can give less weight to common strategies so that, in the reweighted data, all strategies have the same weight: as many trailers go negative as stay positive. Thus, in the reweighted data, the action sequences are balanced across time-varying confounders and there is no omitted variable bias. Crucially, Robins (1999) shows that IPTW estimators do not introduce post-treatment bias because they avoid stratifying the outcome by time-varying confounders.

Alternatives to IPTW estimators for dynamic causal inference include structural nested models, structural equation modeling, synthetic control methods, and principal stratification, but each of these methods has a disadvantage when compared to the weighting approach. The IPTW estimator described below is far more general than the latter three approaches, while being less model dependent than structural nested models. Robins (2000) points out that while these structural nested models can estimate more flexible causal quantities of interest than IPTW, they also require models for each time-varying confounder in the data.[4] While this weighting approach is less flexible than structural nested models, it is much more flexible than other, related methods. Structural equation modeling

3. IPTW estimators have a long history in statistics, beginning with the Horvitz-Thompson estimator (Horvitz and Thompson 1952), which has been applied to many problems outside of causal inference, including survey sampling. For an introduction to these estimators for causal inference in political science, see Glynn and Quinn (2010).

4. For a more detailed discussion of the advantages and disadvantages of IPTW approaches versus g-estimation and structural nested models see Robins (2000).

requires a constant effects assumption in order to estimate dynamic causal effects.[5] Synthetic control methods for comparative case studies, for example, focus on a single intervention for each unit and thus limit the number of possible estimable quantities (Abadie, Diamond, and Hainmueller 2010). Principal stratification (Frangakis and Rubin 2002) can recover causal effect estimates when a post-treatment variable defines the available sample, such as censoring by death. Frangakis and Rubin (2002) note, however, this approach is more appropriate for non-manipulable post-treatment variables. When the relevant post-treatment variable is manipulable and truly part of the treatment, as is the case here, principal stratification needlessly restricts the quantities of interest under investigation.

1.2 A framework for dynamic causal inference

To show how IPTW can estimate causal effects in a dynamic setting, it is useful to extend the single-shot causal inference framework to explicitly include time.[6] Suppose i indexes the campaign, with $i = 1, \ldots, N$. Let t denote the week of the campaign, taking possible values $1, \ldots, T$, where T is the final week before election day. We refer to $t = 1$ as the "baseline" time period; it is the time period before the campaign begins, assumed to be the first week after the primary. In each period, campaigns can either go negative, denoted $A_{it} = 1$ or remain positive, $A_{it} = 0$.

Campaigns face a rapidly evolving environment. To account for this, let X_{it} represent the characteristics of the campaign in week t that affect the Democrat's decision to go negative in week t. This would include recent polling or Republican negativity in the previous weeks. This definition assumes that the decision to go negative occurs "after" the variables in X_{it}, so that they are pre-action for week t.[7] Instead of containing all variables occurring at time t, the set of covariates describes the information setting for the action decision at time t. Simply put, X_{it} is the most recent set of variables

5. See Glynn (2011) for a description of this problem in the context of mediation in linear structural equation models.

6. The following section rests heavily on the potential outcomes-based model of causal inference championed by Rubin, 1978 and extended to dynamic settings by Robins (1986, 1997).

7. The causal ordering here is notationally arbitrary as its reversal would require only a change in subscript. More crucially, researchers must determine what information is pre- and post-action in a given period for the substantive question at hand.

that could possibly affect A_{it}. [8] The baseline covariates, X_{i1}, include background information that remains static over the course of the study. For campaigns, these could be perceived competitiveness of the election, number of ads shown in the primary, incumbency status, or challenger quality. The choice of relevant covariates of course depends on the outcome, Y, which in this case is the Democratic percent of the two-party vote.

Dynamic settings require references to the *history* of a variable. A history is the set of all instances of that variable up to some point in time. In this example, it may be the sequence of campaign tone or poll results in each week. Underlines indicate the history of a variable, so that $\underline{A}_t$ would be the negativity up through time t: $\underline{A}_t \equiv (A_1, A_2, \ldots, A_t)$. The covariate history, $\underline{X}_t$, is defined similarly. One possible realization of $\underline{A}_t$ is $\underline{a}_t \equiv (a_1, \ldots, a_t)$, where each a_t can take the values 0 or 1. Furthermore, let $\underline{A} = \underline{A}_T$ be the sequence of negativity over the course of the entire campaign. Let $\underline{a}$ be a representative campaign tone history and $\underline{\mathcal{A}}$ as the set of all possible values of $\underline{a}$; that is, all the possible ways a candidate could go negative over the course of the campaign. Let $\underline{X}$, $\underline{x}_t$, and $\underline{x}$ be defined similarly for the covariate history.

Each possible negativity sequence, $\underline{a}$, has an associated potential electoral outcome. Let $Y_i(\underline{a})$ be the Democratic percent of the two-party vote if we forced candidate i to implement the campaign $\underline{a}$. Note that there are 2^T possible sequences $\underline{a}$. As before, any individual candidate can experience at most one of these potential outcomes, which is the one associated with the their observed action history. The rest of the potential outcomes will be counterfactual; they are *what would have happened* if the unit had followed a different sequence. Suppose campaigns only lasted two weeks. In this world, $Y_i(0, 1)$ would be the Democratic vote-share if candidate i were to remain positive in week one and go negative in week two. To complete the definition of the potential outcomes, we connect them to the observed outcomes, Y_i. When some unit is observed to have followed action sequence $\underline{a}$, then we observe the potential outcome for that sequence, or $Y_i(\underline{a}) = Y_i$ when $\underline{A}_i = \underline{a}$.[9]

8. Note that these variables are possibly affected by past treatment, but I suppress the potential outcome notation for the covariates for exposition as it adds no consequences for the present discussion.

9. This is commonly referred to as the consistency assumption in the epidemiology literature (Robins, 1997). It implicitly assumes what Rubin (1978) refers to this as the stable unit treatment value assumption or SUTVA. Recent work has further

We say that $\underline{X}_t$ contains a time-varying confounder if it (a) affects the election outcome, (b) affects future negativity, and (c) is affected by past negativity. In estimating the effect of Democrats going negative, the advertising tone of the Republican would be a time-dependent confounder. Democrats are more likely to go negative if their opponent has gone negative and their opponent's actions are likely related to the outcome. Note that X_t could include past values of Y, in which case the lagged dependent variable would be a time-dependent confounder.

1.2.1 The assumptions

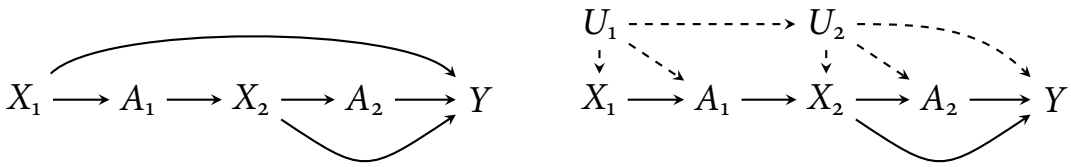
The goal of dynamic causal inference is to estimate the means of the potential outcomes under various action sequences. These are population-based quantities of interest: what would happen if *every* Democrat remained positive? In the sample, however, the candidates that actually went negative always might be different than those who did not. Thus, the sample of units who followed the strategy would be an unrepresentative sample of the potential outcome under that strategy. In order to rid our analysis of the above selection problems, we must be able to identify and measure all possible confounders.

Assumption 1 (Sequential Ignorability). *For any action sequences* $\underline{a}$*, covariate history* $\underline{X}_t$*, and time* t*, if* $\underline{A}_{t-1} = \underline{a}_{t-1}$*, then* $Y(\underline{a}) \perp\!\!\!\perp A_t | \underline{X}_t, \underline{A}_{t-1} = \underline{a}_{t-1}$.

Here, $B \perp\!\!\!\perp C|D$ means that B is independent of C, conditional on D. The assumption of sequential ignorability extends the conditional ignorability assumption to time-varying actions. It states that action decision at time t is independent of the potential outcomes, conditional on the covariate and action histories up to that point. That is, conditional on the past, those who go negative are similar to those who stay positive. Figure 1.2a shows a causal directed acyclic graph (DAG) in which sequential ignorability holds, while Figure 1.2b shows a situation where the assumption fails to hold due to an omitted variable U. If decisions are made by a coin flip, then clearly this assumption will hold. If units act based on the covariate history, however, then it will fail to hold unless the analyst can observe all of

formalized and clarified this assumption (Cole and Frangakis, 2009; VanderWeele, 2009; Pearl, 2010).

those covariates. For instance, the assumption would be violated if campaigns made the decision to go negative based on polling data, but the analyst did not have access to that polling data. The goal for researchers, then, is to collect all the covariates that might influence the decision to go negative in some week. While this is a daunting task in an observational study, it is no harder than satisfying conditional ignorability in the single-shot case and Section 1.5.1 shows how to relax the assumption in a sensitivity analysis.



(a) Sequential ignorability holds. **(b)** Sequential ignorability fails.

Figure 1.2: Directed Acyclic Graphs representing different assumptions about sequential ignorability, where U is an unobserved variable.

Finally, in order to compare the various action sequences, each must have some positive probability of occurring. It is nonsensical to estimate the effect of a sequence that could never occur.

Assumption 2 (Positivity). *For any sequences* $\underline{a}_t = (\underline{a}_{t-1}, a_t)$ *and* $\underline{x}_t$*, and time t, if* $\Pr\left(\underline{A}_{t-1} = \underline{a}_{t-1}, \underline{X}_t = \underline{x}_t\right) > 0$*, then* $\Pr\left(A_t = a_t | \underline{X}_t = \underline{x}_t, \underline{A}_{t-1} = \underline{a}_{t-1}\right) > 0$.

This assumption outlines the types of strategies we can study. Positivity can break down when some sequences fail to occur in the actual data even though they are theoretically possible. In negative advertising, for instance, candidates with extremely safe seats never go negative, even though nothing is stopping them from doing so. Unfortunately, we will be unable to estimate the effect of going negative for these candidates. These empirical violations of positivity are closely related to the assumption of *common support* often invoked in the matching literature. Section 1.5 discusses these practical problems with positivity and how to restrict the analysis to the common support.

1.2.2 Marginal structural models for the potential outcomes

In the single-shot approach, estimating a causal effect only involves two quantities, one corresponding to each action: $E\big[Y_i(1)\big]$ and $E\big[Y_i(0)\big]$. In dynamic causal inference, there is one potential outcome for each action sequence. A key consequence is that even with a small number of time periods, there will be an overwhelming number of possible action sequences. With two potential outcomes, we can non-parametrically estimate the mean outcome in the treated and control groups by taking sample means. With just ten periods, however, there would be 1,024 possible action sequences, making it unlikely that there will be even one unit following any particular sequence. Thus, the non-parametric approach of single-shot methods will be useless here.

To overcome this curse of dimensionality, we can use a parametric model to relate the action sequences to the potential outcomes. That is, we will suppose that "similar" action sequences should have "similar" potential outcomes. Imposing this structure on the problem reduces the dimensionality of the problem at the expense of possible model misspecification. Robins, Hernán, and Brumback (2000) introduced a parsimonious class of semi-parametric models for this problem called marginal structural models (MSM). In this class of models, we assume a parametric form for the mean of the potential outcome

$$\mathrm{E}[Y(\underline{a})] = g(\underline{a}; \beta), \tag{1.1}$$

while leaving the rest of the distribution of $Y(\underline{a})$ unspecified. This model is semi-parametric in the sense that it leaves unrestricted the relationship between the outcome and the covariates, though the implementation of this models trades this for modeling the relationship between the actions and the covariates.

The function g defines our assumptions about which action sequences should have similar potential

outcomes. We may have, for instance,

$$g(\underline{a}; \beta) = \beta_0 + \beta_1 c(\underline{a}), \tag{1.2}$$

where $c(\underline{a}) = \Sigma_{t=0}^{T} a_t$ is the cumulative action. This model assumes that units with the same number of total periods acted should have similar potential outcomes, with β_1 as the causal effect of an additional period of the action. In the context of negative campaigning, β_1 is the effect of an additional week of negativity. An assumption here is that going negative for the first five weeks of the campaign is the same as going negative for the last five weeks of the campaign. Depending on the application, this might be a more or less plausible assumption and, in general, these types of modeling assumptions will always produce some amount of bias. The greater flexibility we allow for $g(\underline{a}; \beta)$, however, the more variable our estimates become. The substance of the problem and the amount of data on hand will determine what model makes sense for the potential outcomes.

Supposing that (1.2) was the correct model for the potential outcomes, we want to estimate its causal parameters. One approach would be to estimate

$$\mathrm{E}[Y|\underline{A} = \underline{a}] = \gamma_0 + \gamma_1 c(\underline{a}), \tag{1.3}$$

which omits any covariates X_t and simply regresses the outcome on the observed action. This approach replaces the potential outcomes $Y(\underline{a})$ with the observed outcomes Y, holding the model fixed. If X_t affects the action and the outcome, however, the associational parameter, γ_1, will not equal the causal parameter, β_1, due to omitted variable bias. That is, differences in the observed outcomes could be due to difference in the covariate history, not the action sequence. We could instead condition on X_t by estimating

$$\mathrm{E}[Y|\underline{A} = \underline{a}, X_t] = \delta_0 + \delta_1 c(\underline{a}) + \delta_2 X_t, \tag{1.4}$$

either through a regression that includes X_t or a matching algorithm which matches on X_t. The key

parameter, δ_1, will still fail to equal the causal parameter of interest, β_1, when X_t is a time-varying confounder, since X_t is post-treatment for $\underline{A}_{t-1}$. Thus, X_t is in the difficult position of being both an omitted variable *and* a post-treatment variable for the action history. These traditional methods of estimating β_1 fail in the face of time-varying confounders, whether or not we adjust for them, since either approach leads to bias.

One might think that the two traditional estimation procedures would at least provide bounds on the true causal effect, with β_1 falling between γ_1 and δ_1. When the omitted variable bias and the post-treatment bias have the same sign, however, this bounding will fail to hold. This can occur, for instance, when strategic actors attempt to compensate poor performance with beneficial actions. Suppose that there is a strong, positive effect of negative advertising and that trailing campaigns use it to bolster their positions. The omission of polling in a model would lead to an understatement of the negativity effect, since candidates tend to be trailing when they go negative. Positive campaigns would appear stronger than negative campaigns, even though negativity boosts performance. The inclusion of polling in a model would also lead to an understatement of the effect, since it washes away the increase in polls from past negativity. Thus, the true effect of negativity would be higher than either of the traditional methods would predict. Robins (1997) gives a numerical example that has these features.

1.3 The action model and inverse-probability of treatment weighting

As shown above, the usual single-shot approaches break down when the actions can vary over time. Fortunately, inverse-probability of treatment weighting (IPTW) can recover unbiased estimates of causal effects, even in dynamic settings. To see how IPTW works, note that, due to the omitted variables, the distribution of the potential outcomes differs from the distribution of the observed outcomes ($E[Y(\underline{a})] \neq E[Y|\underline{A} = \underline{a}]$). Regression and matching attempt to avoid this problem by finding subsets of the data where those distribution are the same and making comparisons within these subsets. This conditioning removes the omitted variable bias, but it can induce post-treatment bias. Methods that

rely on weighting, such as IPTW, avoid these by never explicitly conditioning on the confounders in the outcome model.[10]

Robins, Hernán, and Brumback (2000) show that under the above assumptions, a reweighted version of the observed outcomes will have the same distribution as the potential outcomes. In the campaigns context, the reweighted outcomes for always-positive campaigns will look like the outcomes if we forced all Democrats to remain positive. The weights in a given week are defined as

$$W_{it} = \frac{1}{\Pr(A_{it}|\underline{A}_{it-1}, \underline{X}_{it})}. \tag{1.5}$$

In words, the denominator of W_{it} is the probability of observing the action sequence that unit i actually took in that week, conditional on the past. To generate an overall weight for each race, we simply take the product of the weekly weights over time:

$$W_i = \prod_{t=1}^{T} W_{it}. \tag{1.6}$$

A simple example helps to explain the construction of the weights. Suppose that there were only two weeks in a campaign, with a poll update in between the weeks. A candidate decides to go negative or stay positive in the first week, sees the outcome of the poll, decides to go negative in the second week, and then observes the election results. A candidate who stays positive in week one, trails in the polls, and then goes negative in week two would have the following weight:

$$W_i = \frac{1}{\Pr(\text{pos}_1)} \cdot \frac{1}{\Pr(\text{neg}_2|\text{trail}, \text{pos}_1)}. \tag{1.7}$$

10. This approach is similar in spirit to Heckman selection models (Heckman 1976; Achen 1986) in the sense that they separate out the selection model (who goes negative) and the outcome model (how negativity affects vote shares). These types of methods, however, rely on instrumental variables. Unfortunately, instrumental variable methods require effects to be constant in the population in order to estimate the average causal effect (Angrist, Imbens, and Rubin 1996). The approach described here which make no such assumptions.

The first term in the denominator is simply the probability of being positive in the first week. The second term is the probability she would have gone negative in the second week, conditional on trailing and having been positive in the first week. The resulting denominator is the probability of observing the campaign $(\text{pos}_1, \text{neg}_2)$, conditional on the time-varying covariate, polls.

1.3.1 Why weighting works

The weights in IPTW remove any confounding by ensuring that the distribution of action sequences is the same in each level of the confounder. In the reweighted data, the action decisions are unrelated to the measured confounders and, thus, they cannot account for any remaining differences between action sequences. It is instructive to see how this works in the single-shot case. Let $\Pr_W(\text{neg}|\text{trail})$ be the reweighted probability of observing an a negative candidate, conditional on trailing in the polls. We can find this probability by multiplying the original probability by the weight for this type of observation and divide by a normalizing constant:

$$\Pr_W(\text{neg}|\text{trail}) = \frac{\frac{1}{\Pr(\text{neg}|\text{trail})} \cdot \Pr(\text{neg}|\text{trail})}{\frac{1}{\Pr(\text{neg}|\text{trail})} \cdot \Pr(\text{neg}|\text{trail}) + \frac{1}{\Pr(\text{pos}|\text{trail})} \cdot \Pr(\text{pos}|\text{trail})} = \frac{1}{2}. \tag{1.8}$$

The denominator simply ensures that the reweighted probabilities will sum to one. Using the same logic, it is clear to see that $\Pr_W(\text{neg}|\text{trail}) = 1/2$ as well. Thus, in the reweighted data, a race is equally likely to go negative as stay positive when they are trailing.

Intuitively, this weighting breaks the links between the action decision and the factors that affect the action decision. Candidates that are pursuing common strategies, where $\Pr(A_{it}|\underline{A}_{it-1}, X_{it})$ is closer to 1, will have lower weights than those candidates with less common strategies. This weighting corrects the deviations from the ideal experiment we would have like to run. The sequential ignorability assumption is crucial here because we cannot correct for deviations we do not observe. Since there is no connection between the action sequence and the confounders in the reweighted data, we can simply run whatever model we wanted to run in the first place, without conditioning on time-varying confounders. And

because we never condition on these variables, we never introduce post-treatment bias as with single-shot approaches.

1.3.2 Estimating the weights

Of course, without randomization, the probability of going negative will be unknown, leaving (1.5) to be estimated. To do so, we must model the decision to go negative in each week, conditional on the past. Since the decision is dichotomous, a common approach is to estimate the probability of going negative with a logit model:

$$\Pr(A_{it} = 1 | \underline{A}_{it-1}, \underline{X}_{it-1}; \alpha) = \left(1 + \exp\left\{-h\left(\underline{A}_{it-1}, \underline{X}_{it}; \alpha\right)\right\}\right)^{-1}, \tag{1.9}$$

where h is a linear, additive function of the action history, covariate history, and parameters α. For instance, we might have

$$h\left(\underline{A}_{it-1}, \underline{X}_{it}; \alpha\right) = \alpha_0 + \alpha_1 A_{it-1} + \alpha_2 X_{it} + \alpha_3 t, \tag{1.10}$$

which models the action decision as a function of negativity in the last week (A_{t-1}), the most recent poll results (X_{it}), and the week of the campaign (t).

An estimate of the weights requires an estimate of the parameter vector α from this model. We can obtain these estimates, $\hat{\alpha}$, from a pooled logistic regression, treating each campaign-week as a separate unit. These estimates form the basis for the estimated weights,

$$\widehat{W_i} = \prod_{t=1}^{T} \frac{1}{\Pr\left(A_{it} | \underline{A}_{it-1}, \underline{X}_{it}; \hat{\alpha}\right)}, \tag{1.11}$$

where the denominator is now the predicted action probability (or fitted value) for unit i at each time period.[11] Note that it is not necessary to estimate the same model for all units. For example, incumbents

11. These values are easily found using a combination of the `glm` and `predict` functions in R (R Development Core Team,

and non-incumbents might require different models because their approaches to the negativity decision are so distinct.

Each observation i is then weighted by $\widehat{W_i}$ in a weighted generalized linear model for the outcome, with form $g(\underline{a}; \beta)$ from (1.2).[12] Robins (2000) shows this estimation procedure is consistent for the causal parameters, β, under sequential ignorability, positivity, and the correct model for the weights. The most straightforward way to estimate standard errors and confidence intervals is to bootstrap the entire estimation procedure, including the weights (Robins, Hernán, and Brumback, 2000). For negative campaigning, this means resampling the set of campaigns (not the set of campaign-weeks), re-estimating the weights, and running the weighted outcome model on the resampled data.

1.3.3 Stabilized Weights

If campaigns have vastly different likelihoods of going negative, then the estimated weights from (1.11) can have extreme variability, which results in low efficiency. We can use a slightly different version of the weights, called the *stabilized weights*, to decrease this variability and increase efficiency. The stabilized weights take advantage of an interesting fact: the numerator of the weights does not change the consistency of the estimation procedure.[13] While it was natural to use the value 1 as the numerator, we can replace it with other functions of the action history that increase efficiency. The usual choice used in the literature () is

$$SW_i = \prod_{t=1}^{T} \frac{\Pr\left(A_{it} | \underline{A}_{it-1}; \delta\right)}{\Pr\left(A_{it} | \underline{A}_{it-1}, \underline{X}_{it}; \alpha\right)}, \tag{1.12}$$

where the numerator is a model for the marginal probability of action, conditional on past action. When actions are randomized, these stabilized weights will be equal to one since the action probability would be unaffected by the covariates in the denominator.

2011).

12. The `survey` package in R can implement this weighting for a large class of outcome models (Lumley, 2004).

13. As shown in Robins (2000), the numerator only alters the marginal distribution of the action $\underline{A}$, which does not affect the marginal distribution of the potential outcomes, $Y(\underline{A})$. This is because the marginal distribution of $\underline{A}$ does not affect the distribution of Y conditional on $\underline{A}$, which is the crucial ingredient for the distribution of the potential outcomes.

Of course, the numerator of SW_i is unknown, leaving us with the task of estimating δ. All this requires is an additional logit model for the numerator to estimate the probability of going negative without conditioning on the time-varying covariates. If the outcome model will include interactions with baseline covariates, then both the numerator and the denominator should include those variables. To construct these weights, one simply needs to obtain predicted probabilities from each model for every unit-period. Then, for each unit, take the product of those probabilities across time periods and divide to obtain the estimates $\widehat{SW_i}$.

1.4 Estimating the effect of going negative

Pundits and theorists often bemoan the growth in negative campaign advertising in recent decades. Less often do they discuss its effectiveness. An implicit assumption in the air of political discourse is "Of course it works, politicians do it." The prospect of dirtying the waters with such cheap and tawdry tactics is bad enough, being useless would only add insult to injury. A contingent of political scientists have investigated just how useful negativity is for candidates, without reaching a consensus. In a comprehensive review, Lau, Sigelman, and Rovner (2007) describe the state of the literature: "All told, the research literature does not bear out the idea that negative campaigning is an effective means of winning votes."

The usual approach to estimating the effectiveness of negative advertising (see Lau and Pomper (2002, 2004) for examples) is to regress election outcomes on a summary measure of the degree of negativity in a campaign along with controls for various static attributes of the race. A crucial problem for these investigations is that campaign tone is a dynamic process, changing from week to week. Furthermore, there are strong time-varying confounders. For instance, poll numbers affect the decision to go negative, but going negative also affects poll numbers. Thus, polling is both pre- and post-action: a classic time-varying confounder. As shown above, ignoring the polls and conditioning on the polls will both result in biased estimates. We can estimate the effect of time-varying actions, though, using

marginal structural models and inverse probability of treatment weighting.

The goal of this application is to estimate the effect of going negative for Democratic candidates in state-wide elections. I use data on campaigns for Senate and Gubernatorial seats in the cycles of 2000, 2002, 2004, and 2006. For each campaign, I code the advertising tone using data from the University of Wisconsin Advertising Project (Goldstein and Rivlin 2007). To ensure consistency across years, I use a simple measure of negative or contrast ads: does the ad mention the opposing candidate?[14] I use this coding to construct a measure of whether a candidate has "gone negative" in a given week of the campaign based on what percentage of ads are negative.[15] The WiscAds data also provide a proxy for weekly campaign spending: the total number of ads aired in a week. In addition to advertising data, I also collected weekly polling data from various sources,[16] along with baseline covariates, such as predicted competitiveness of the race (as measured by the *Congressional Quarterly* score), incumbency status, number of ads run by each candidate in their primaries, the length in weeks of the campaign, measures of challenger quality and incumbent weakness, and the number of Congressional districts in the state. Much of this data comes from Lau and Pomper (2004) with additional data collection. In this example, baseline is the day after the final primary.

1.4.1 A model for going negative

In order to estimate the causal parameters from an MSM, we must construct the weights from Section 1.3.3. In order to satisfy the assumption of sequential ignorability, we must gather as many covariates as possible that might influence the decision to go negative in a given week and are correlated with the election outcome. This is, of course, a difficult task, but we can often leverage substantive knowledge to guide our models. The dynamic reasons for a candidate to go negative might be numerous, but it is

14. The WiscAds project failed to collect data in 2006, so I acquired and computer-coded the data directly from CMAG, the consultant group which provides the data to WiscAds.

15. For the analysis below, I used a cutoff of 10%. The results appear unaffected by this choice, as weeks tend to be dominated by only a few ads. If there is one ad that is negative, it pushes the percent negativity quite high.

16. Polling data comes from the *Hotline* daily political briefing for 2000 and 2002 and from `http://www.pollster.com` for 2004 and 2006.

likely that the state of the race, as summarized by the polls in a given week, are at worst a proxy for these factors and at best the most important factor. Indeed, we might think that the candidates doing the worst in the polls are the most likely to go negative. This is why it is crucial to include both pre-campaign measures of competitiveness and dynamic measures of campaign performance in the form of polling data. Without this data, it would be impossible to differentiate between the effect of negative advertising on the one hand, and negative advertising simply indicating weak candidates on the other. Previous literature on the effectiveness of negative advertising has not had access to the kind of polling data available in more recent elections, hampering their ability to address these dynamic selection issues.

To address these concerns I include the following time-varying covariates in the weighting model: the Democratic share of the polls in the last week, the share of undecided voters in the last week, past negativity of both Democrats and Republicans, the amount of advertising by both candidates in the last week, and the week of the campaign. It may be the case that these variables do not encompass or proxy all of the factors that influence candidates, which is why it is crucial to assess any inferences using sensitivity analysis, as I do in Section 1.5.1. With these covariates, I ran two separate pooled-logistic models for the decision to go negative: a separate numerator and denominator model.[17]

These models largely fit with the intuition and theory of campaigns, with high-advertising and already-negative races being more likely to be negative. Figure 1.3 shows that there is a strong relationship between polling and the decision to go negative: non-incumbent Democrats in safe seats rarely go negative, but those who are trailing often do. To construct the weights, I combine predicted probabilities from these models according to (1.12).[18] Due to empirical violations of positivity, I restricted the analysis to common support on baseline covariates, which mostly involved removing

17. In order to stabilize the weights further, I include all baseline covariates in both the numerator and denominator models. This means that IPTW will only balance the time-varying covariates, leaving any remaining baseline imbalance. Since this imbalance is time-constant, we can remove it through traditional modeling approaches and, thus, I include those covariates in the outcome model below.

18. The weight models are pooled logistic generalized additive models (GAMs), which is what allows for the flexible modeling of the polling. I used the `mgcv` package to fit this model (Wood, 2011).

extremely uncompetitive race.[19]

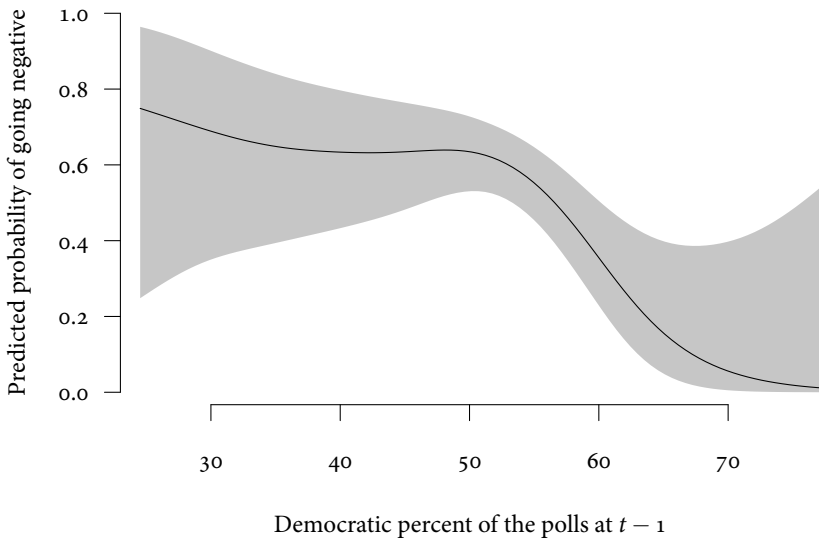


Figure 1.3: The marginal relationship between lagged polling numbers and going negative for Democratic non-incumbent candidates. All other variables from the model held at their mean, or median, depending on the type of variable. The shaded region is a 95% confidence bands. Intuitively, trailing Democrats are more likely to go negative than leading Democrats.

1.4.2 The time-varying effects of negativity

The effect of negative advertising is unlikely to be constant across time. Ads closer to election day should have a stronger impact than those earlier in the campaign and marginal structural models allow us to estimate these time-varying effects. I break up the effect into a an early campaign effect (the primary through September) and a late campaign effect (October and November). Vote shares are

19. Note that weeks in which a candidate runs *no* ads is week where a candidate *cannot* go negative. These weeks receive a stabilized weight of one, meaning they do not contribute to the weight of their overall campaign. A more thorough analysis would treat the number of ads and the tone of those ads a joint treatment.

Estimator	Democratic Incumbent	Democratic Non-incumbent
Naïve	-0.96 (-1.68, -0.33)	0.49 (-0.18, 1.20)
Control	-0.54 (-1.28, 0.11)	0.63 (-0.02, 1.26)
IPTW	-0.70 (-1.47, 0.01)	0.75 (0.21, 1.27)

Table 1.1: Estimated effects of an additional week of negative advertising in the last five weeks of the campaign on the Democratic percent of the two-party vote. Bootstrapped 95% confidence intervals are in parentheses, with those crossing zero set in gray. Inverse probability weighting estimates a strong, positive effect for non-incumbents and a strong, negative effect for incumbents. Note that the competing models fail to bound the IPTW-estimated effect.

continuous, so a linear MSM is appropriate for the potential outcomes:

$$E[Y_i(\underline{a})] = \beta_0 + \beta_1 \left(\sum_{T-5}^{T} A_{it} \right) + \beta_2 Z_i + \beta_3 Z_i \left(\sum_{T-5}^{T} A_{it} \right) + \beta_4 \underline{A}_{iT-6} + \beta_5 Z_i \underline{A}_{iT-6} + \beta_6 X_i, \quad (1.13)$$

where Z_i is an indicator for being a Democrat incumbent, X_i is a vector of baseline covariates, and T is the week of the election. The summation terms calculate how many of the last five weeks of the campaign the Democrat went negative. This covers October and early November, which is the home-stretch of the campaign. The model separately estimates the direct effect of earlier negativity and allows for incumbent status to modify both early and late effects. Following the IPTW approach, I weight each campaign using weights constructed from the above "going negative" model.

It is instructive to compare estimates from this model with two competing approaches. First, the *naïve estimator* simply ignores all time-varying covariates and fits (1.13) to the observed data without weights. Second, the *control estimator* attempts to control for the covariates by including them as additional regressors in (1.13).[20] These represent the two single-shot methods used by applied researchers: the naïve estimator to guard against post-treatment bias and the control estimator to guard against omitted variable bias.

Table 1.1 shows the estimated effects of late campaign negativity from all three models broken out by

20. Polls are included as the mean Democrat poll percentage, total number of ads as the average ads per week, and Republican negativity as the overall duration of Republican negativity.

incumbent status.[21] The MSM finds that Democratic incumbents are hurt by going negative, while non-incumbents are helped. Non-incumbents see a 0.72 percentage point increase in the Democratic percent of the two-party vote for every additional week of negative advertising in the last five weeks. Incumbents, on the other hand, drop 0.70 percentage points for the same change. As Figure 1.4 shows, there is no evidence of a direct effect of earlier negativity on the final vote in either group. Note that these results control for polls taken at the beginning of the campaign. It is surprising, then, to see effects that are even this large since these baseline polls are highly predictive of the outcome in Senate and Gubernatorial elections.

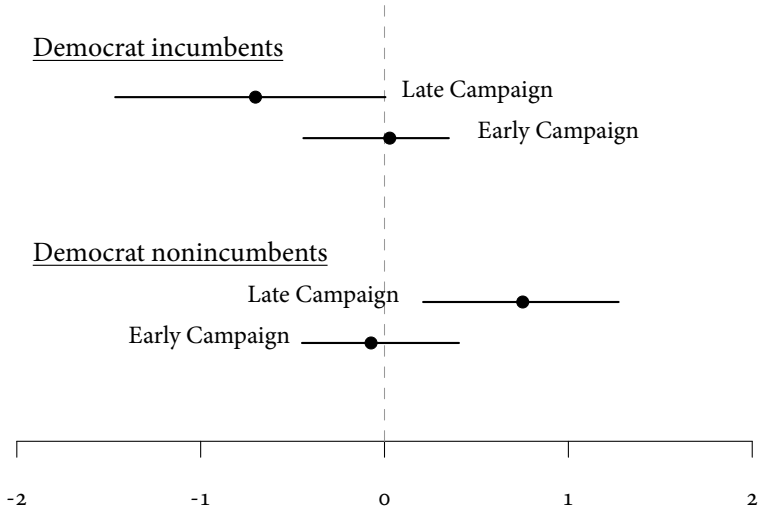


Figure 1.4: Inverse-probability of treatment weighting estimates of the time-varying effects of negative campaigning with bootstrapped 95% confidence intervals. Negative ads are more potent later in the campaign (October and November) than earlier in the campaign, but the direction of the effect is negative for incumbents and positive for non-incumbents.

21. Estimates here are produced using the `svyglm` function in the `survey` package, version 3.22-4 (Lumley, 2010). Standard errors and confidence intervals come from bootstrapping this model.

When we compare the three estimators, we find that using a dynamic causal inference approach leads to substantively different conclusions about negative advertising. For non-incumbents, for instance, the two single-shot methods recover what the previously literature has found: no significant effect of negativity (Lau and Pomper, 2004; Lau, Sigelman, and Rovner, 2007). With the MSM, however, we find that negativity does have a large, statistically significant, and positive effect. Interestingly, the MSM-estimated effect is well outside of the bounds set by the naïve and control estimators. For these non-incumbents, the IPTW estimate is over 18% larger in magnitude than either of the other methods. Thus, "trying it both ways" would be an unsuccessful strategy in this case. For incumbents, we find additional divergence from the literature. Similar to the effects of negativity for incumbents estimated by (Lau and Pomper, 2004), the naïve estimator finds a large and harmful effect of negativity for incumbents. But this is likely driven by mid-campaign confounding and once I account for this using a MSM, I find no significant effect of negativity for incumbents. Overall, a dynamic causal inference approach leads to different conclusions about the effectiveness of negative advertising for U.S. statewide election than the previous literature had found.

1.5 Assessing model assumptions

With single-shot causal inference methods such as matching, balance checks are crucial diagnostics (Ho et al., 2006). These checks ensure that the treated and control groups are similar on their background covariates. Usually this takes the form of simple comparisons of covariate means in the treated and control group, though more sophisticated techniques exist. Unfortunately, this simple approach is ill-suited to the dynamic setting since it is unclear what groups to compare. At a given week of the campaign, negative and positive campaigns might differ on a time-varying confounder, but these differences might be due to past negativity.

Under the above assumptions of the IPTW estimator, the decision to go negative is unconfounded in the weighted data, conditional on past negativity. We should expect, then, that the observed actions will

be independent of time-varying covariates once we weight by SW_i. This independence is, however, conditional on a unit's action history. For instance, suppose we had two campaigns that had remained positive until week t. Then the decision to go negative in week $t+1$ for these two campaigns should not depend on time-varying covariates, such as polling, in the weighted data. We can assess balance in the weighted data, then, by checking for associations between the action decision and the time-varying covariates that affect that decision, conditional on the action history. If, after reweighting the data and conditioning on past negativity, the decision to go negative is still predictive of past polling, then there is likely residual confounding of the relationship between the outcome and negativity.

Figure 1.5 shows how the weights reduce this *history-adjusted imbalance* in the campaign advertising example. It shows the change in the standardized history-adjusted imbalance from the unweighted to the weighted data.[22] In the unweighted data, for instance, Democrats were much more likely to go negative after an attack by Republicans (R Neg_{t-1}). Once we apply the weights, however, the differences move much closer to zero because IPTW gives relatively more weight to races that went negative without Republican negativity in the last week.

One stark observation from the diagnostic plot is that confounding exists for incumbents even after weighting for the time-varying confounders. This makes sense for two reasons. First, data quality issues plague incumbents since their safer races attract less polling. Second, incumbents have stronger positivity problems with extremely safe seats rarely going negative. Furthermore, incumbent campaign-weeks with high total ad volumes almost always feature negativity. These issues prevent the weights from fully eliminating the confounding in the data and should give us pause when interpreting the estimates for incumbents.

Cole and Hernán (2008) propose a series of models checks based on the distribution of the weights, SW_i. They note that the confounding of time-varying covariates is what pushes weights away from 1. A

22. These differences come from a unweighted and weighted pooled regression of the time-varying covariate at t on (a) the baseline covariates, (b) Democratic negativity before week t, and (c) Democratic negativity in week t. The coefficient on (c), divided by its standard error, is the standardized difference.

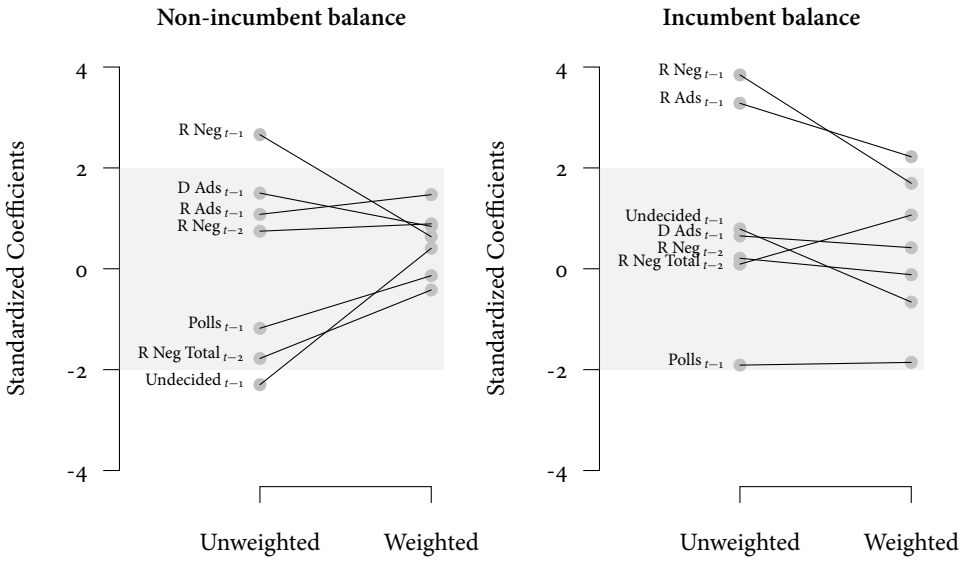


Figure 1.5: The change in history-adjusted balance between the weighted and unweighted data as measured by standardized differences between those campaign-weeks that went negative versus those that remained positive, conditional on baseline covariates. Note that the differences are, all told, closer to zero in the weighted model. "R Neg" is whether the Republican went negative, "Ads" are the number of ads run by the candidates, "R Neg Total" is the total number of Republican negative weeks in the campaign, and "Polls" is the averaged polling numbers for Democrats.

mean weight far below 1 indicates that there are relatively few surprise actions—those that are unlikely given the covariate history. This lack of "surprises" indicates that the probability of going negative is close to 0 and 1 in some parts of the covariate space, which is a violation of positivity. In some sense, there are very few comparable cases for these units with high weights. In the matching literature, this is called a lack of common support. A good check for these issues in the weight model is to check the distribution of stabilized weights period to ensure that (a) the means at each point in time are close to 1 and (b) the minimum and maximum values are reasonably close to 1. The final distributions of the weights by week are in Figure 1.6. Their means are all very close to one and the upper bounds are fairly low, indicating well-behaved weights.

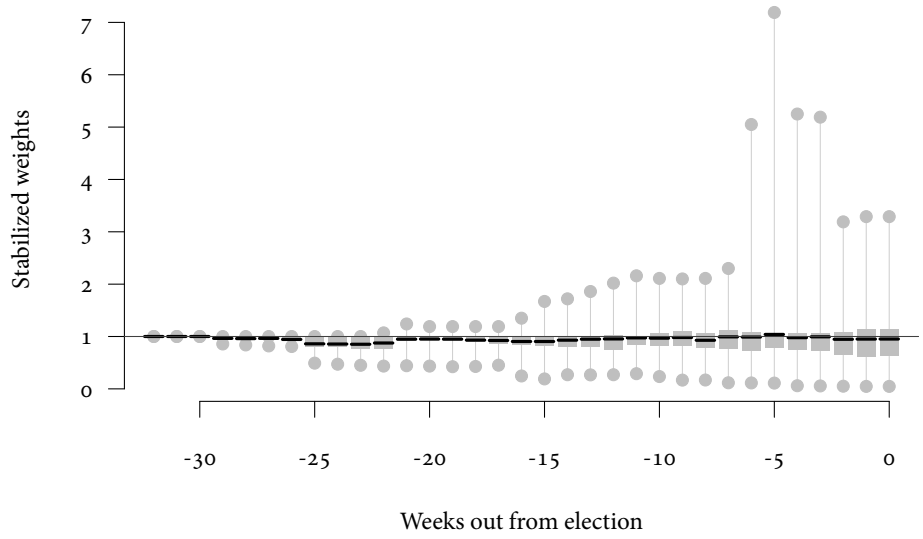


Figure 1.6: Stabilized weights over the course of the campaigns. The black lines are the weekly means, the gray rectangles are the weekly inter-quartile ranges, and the thin gray lines denote the range of the weights. Note that campaigns begin at various times so that there are very few campaigns at 30 weeks out, but very many at 5 weeks out. These weights appear well-behaved as their means are close to 1.

1.5.1 Sensitivity Analyses

Causal estimates from an MSM have excellent properties when the assumptions of positivity and sequential ignorability hold. Of these, sequential ignorability is the trickiest, as it requires that, conditional on the covariate and action histories, the action decision is unrelated to the potential outcome. Any residual differences in potential outcomes between treated and control groups we call *unmeasured confounding* or omitted variable bias. Unless we conduct an experiment and randomize the action, this assumption must be justified through substantive knowledge. Since it is impossible to test this assumption, it is vital to include as much information as possible and to conduct a sensitivity analysis of any estimated results.

Robins (1999) proposes a method to investigate the sensitivity of estimates to the presence of

unmeasured confounding. Robins quantifies the amount of confounding as

$$q(\underline{x}_t, \underline{a}_{t-1}, a_t^*) = E[Y(\underline{a})|\underline{x}_t, \underline{a}_{t-1}, a_t] - E[Y(\underline{a})|\underline{x}_t, \underline{a}_{t-1}, a_t^*]. \tag{1.14}$$

This function measures the difference in potential outcomes for some group that shares an action and covariate history. For two campaigns that are observational equivalent up to week t, q measures the structural advantages of those campaigns that go negative in week t compared to those that remain positive in week t. For instance, the confounding may take the form

$$q(\underline{x}_t, \underline{a}_{t-1}, a_t^*; \alpha) = \alpha[a_t - a_t^*]. \tag{1.15}$$

This is a simple and symmetric form of omitted variable bias. When α equals zero, then there is no difference between those campaigns that go negative in week t and those that stay positive, given campaign histories. If α is positive, then negative campaigns are instrinsically better than those that remain positive. That is, when α is positive, then $Y(\underline{a})$ is higher for $a_t = 1$ (negative campaign-weeks) than $a_t = 0$ (positive campaigns-weeks). If α is negative, then those candidates who are going negative are worse off. Note that these selection biases are all conditional on the observed covariate history.

The above IPTW estimation procedure assumes that $\alpha = 0$, yet Robins (1999) shows that we can estimate the parameters under any assumption about α. Thus, by setting α to various levels, we can estimate the causal effect under different assumptions about the degree of omitted variable bias. To do so, we have to replace the outcome with a *bias-adjusted* outcome,

$$Y_\alpha \equiv Y - \sum_{k \in (0,1)} \sum_{t=0}^{T} \underbrace{q(\underline{X}_t, \underline{A}_{t-1}, k; \alpha)}_{\text{bias at this history}} \cdot \underbrace{\Pr(A_t = k|\underline{A}_{t-1}, \underline{X}_t)}_{\text{probability of reaching this history}}, \tag{1.16}$$

where the first term is simply the observed outcome, Y, and the second term is the overall omitted variable bias, built up from the bias in each time period. We can then re-estimate the parameters of the

marginal structural model with outcome Y_α instead of Y to get bias-adjusted estimates and bias-adjusted confidence intervals. Note that when $\alpha = 0$, the bias function is zero, so that $Y_0 = Y$ and the usual estimation aligns with the assumption of sequential ignorability. Of course, the probability term is unknown and must be estimated. Fortunately, we have already estimated this function as part of the estimation of the weights, SW_i.

Figure 1.7 shows how the estimated effect of late-campaign negativity varies across different assumptions about the omitted variable bias, encoded in the parameter α, which runs along the x-axis. The magnitude of α describes how much stronger or weaker the negative campaigns are, on average, in terms of their potential outcomes. This figure also charts the change in the confidence intervals under the various assumptions about bias, with those that cross zero shaded lighter.

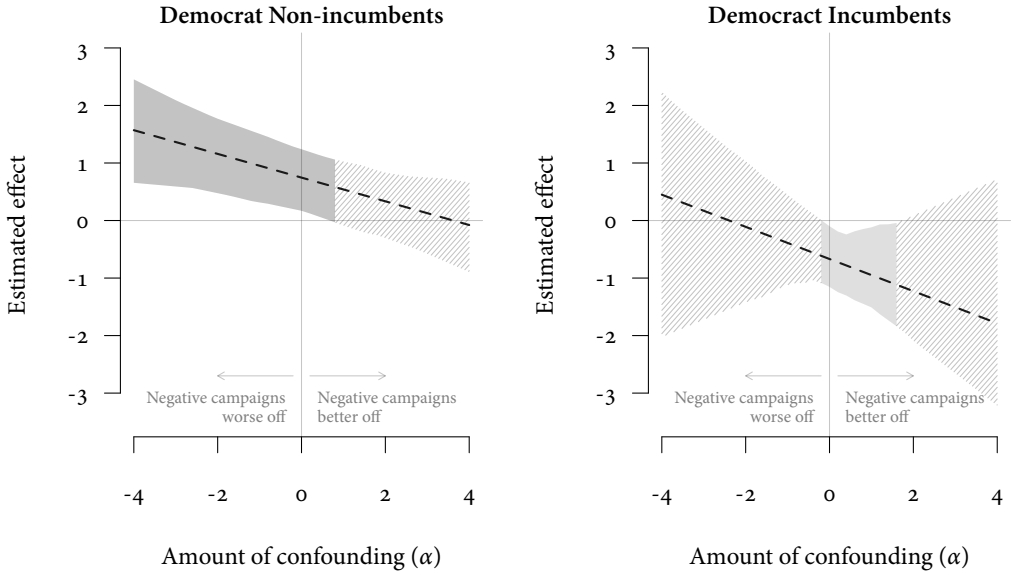


Figure 1.7: Sensitivity of the results to deviations from the sequential ignorability assumption. The parameter α indexes assumptions about confounding, where negative values indicate that the observed negative campaigns are inherently weaker than predicted by the observed variables. Positive values assume that those negative campaigns are stronger than predicted.

When negative campaigns are even 0.5 percentage points stronger than positive campaigns on

average, our 95% confidence intervals would overlap with zero for non-incumbents. This might occur if campaigns were attacking their opponents for (unmeasured) behavior such as a scandal or unpopular vote. Note that these imbalances would have to occur within levels of the covariate history and thus exist after conditioning on polls. If the negative campaigns are instead weaker, perhaps because campaigns go negative when they are in trouble, then results only grow stronger for non-incumbents. The results for incumbents highlight the potential violations of ignorability for that group. The results are fairly sensitive to the degree of confounding, both in the point estimates and the confidence intervals. Note, though, that this confounding would have be above and beyond any information contained in polls and pre-campaign measures of competitiveness.

It is clear that there is some sensitivity to the sequential ignorability assumption and we could attempt to gather more data (quantitative and qualitative) to justify which direction this confounding is likely to lean. Notably, we could measure a larger set of dynamic campaign features such as scandals and endorsements. It would also help to draw on electoral cycles, such as 2008, when incumbents faced strong challengers to overcome the overlap and ignorability issues.

1.6 Conclusion

Political actions do not happen all at once. There are sequences of events that unfold over time. As we have seen, this poses strong problems for extant single-shot causal inference methods. This essay brings to bear a framework that explicitly models the dynamic sequences and builds methods to test their effects. The original application of MSMs was to epidemiological data. Robins (1997) develops a set of methods called structural nested models with an application to HIV treatment studies. In that context, the units are patients and doctors change the treatment over time if the patient status worsens. The analogy to politics is suggestive: campaign managers and candidates as doctors, working to save their patient, the election. Of course, candidates face human opposed to viral opponents, yet this changes only the types of variables needed to satisfy sequential ignorability.

The the structural nested model of Robins (1997) provide an alternative approach to dynamic causal inference. These techniques center on modeling the effect of going negative at every possible history which allows effects that interact with time-varying covariates. The estimation methods resemble backwards induction in game theory. Unfortunately, these structural nested models require models for entire set of time-varying covariates and complicated computation to estimate while researchers can easily use off-the-shelf software to implement an MSM.

The focus of this essay has been the effect of action sequences, yet in many political science situations, actors follow dynamic strategies—updating their actions based on changing conditions. It is likely that the optimal action is actually a strategy, since being able to respond to the current state of affairs is more effective than following a predefined sequence of actions. Hernán et al. (2006) demonstrates that marginal structural models and inverse-probability weighting can estimate the effectiveness of strategies with a simple form such as "go negative when polls drop below x%." In addition, structural nested models can estimate the effect of arbitrary strategies. As might be expected, precisely estimating these effects requires larger sample sizes than the effects of simple action sequences.

A crucial path for future research is model development. In this essay, I used a fairly simple model to estimate different effects for early and late in the campaign. This is a crude division of the data and more fine-grained modeling might help to smooth effects over time. Indeed, we would expect that the effect of negativity in week 5 should be quite similar to effect of negativity in week 6. Better MSMs should be able to handle this type of structure.

Dynamic causal inference is a problem for more than just campaigns. Each subfield of political science analyzes actions that occur over time and have multiple decision points: foreign aid, interest rates, budget allocations, state policies, and even democracy. Indeed, many of the assumptions in this essay (or variations thereof) are implicit in time-series cross-sectional TSCS models, where the counterfactual framework is rarely discussed in explicit terms. Thus, there is a great opportunity for future work that identifies areas with dynamic causal inference problems and attempts to clarify or

improve existing results.

2

Multiple Overimputation: Unifying Measurement Error & Missing Data

2.1 Introduction

Social scientists recognize the problem of measurement error in the context of data collection, but seem to ignore it when choosing statistical methods for the subsequent analyses. Some seem to believe that analyses of variables with measurement error will still be correct on average, but

this is untrue; others act as if the attenuation that occurs in simple types of random measurement error with a single explanatory variable holds more generally, but this too is incorrect. More sophisticated application-specific methods for handling measurement error exist, but they are complicated to implement, require difficult-to-satisfy assumptions, and often lead to high levels of model dependence; few such methods apply when error is present in more than one variable or are used widely in applications, despite an active methodological literature. Unfortunately, the corrections used most often are the easiest to implement but typically have the strongest assumptions, which we discuss below (see Stefanski, 2000 and Guolo (2008) for literature reviews). As with missing data problems a decade ago, many current empirical literatures could benefit from a comprehensive, easy-to-use approach.

We address this challenge by offering a unified approach to correcting for problems of measurement error and missing data in a single easy-to-use procedure. We do this by generalizing the multiple imputation (MI) framework designed for missing data (Rubin 1987; King et al. 2001) to broadly deal with measurement error as partially missing information and treat completely missing cell values as an extreme form of measurement error. The proposed generalization, which we call *multiple overimputation* (MO), enables researchers to treat cell values as either observed without (random) error, observed with error, or missing. We accomplish this by constructing prior distributions for individual cells (or entire variables) with means equal to the observed values, if any, and variance for the three data types set to zero, a (chosen or estimated) positive real number, or infinity, respectively.

Like MI, the easy-to-use MO procedure involves two steps. First, analysts use our software to create multiple (usually about five) data sets by drawing them from their posterior predictive distribution conditional on all available observation-level information. This procedure leaves the observed data constant across the data sets, imputes the missing values from their predictive posterior as usual under MI, and *overimputes*, that is, replaces or overwrites the values or variables measured with error with draws from their predictive posterior. Our basic approach to measurement error, which involves relatively minimal assumptions, allows for random measurement error in any number or combination

of variables or cell values in a data set. With somewhat more specific assumptions, we also allow for measurement error that is heteroskedastic or correlated with other variables. As we show, the technique is relatively robust to violations of either set of assumptions and easy to apply.

An especially attractive advantage of MO (like MI) is the second step, which enables analysts to run whatever statistical procedure they would have run on the completed data sets, as if all the data had been correctly observed. A simple procedure is then used to average the results from the separate analyses. The combination of the two steps enables scholars to overimpute their data set once and to then set aside the problems of missing data and measurement error for all subsequent analyses. As a companion to this essay, we have modified a widely used MI software package to also perform MO (Honaker, King, and Blackwell 2010).

Section 2.2 describes our proposed MO framework, in the context of multiple variables measured with random error with a known, assumed, or completely unknown variance. There, we generalize the MI framework, prove that a fast existing algorithm can be used to create imputations for MO, and offer Monte Carlo evidence that it works as designed. Section 2.3 goes further by deriving methods of estimating the (possibly heteroskedastic) measurement error variance so it need not be assumed. Section 2.4 generalizes our approach further still by allowing measurement error that is correlated with the true values of the variables. Section 2.5 then offers empirical illustrations.

2.2 A Multiple Overimputation Model

We conceptualize the linkage between measurement error and missing data in two equivalent ways. In one, measurement error is a specific type of missing data problem where observed proxy variables provide probabilistic prior information about the true unobserved cell values. In the other, missing cell values have an extreme form of measurement error where no available prior information exists. Either way, the two methodological problems go well together because variables (or cell values) measured with error fall logically between the extremes of observed without error and completely unobserved. This

dual conceptualization also means that our MO approach to measurement error has all the advantages of MI in ease of use and robustness (Schafer 1997; Freedman et al., 2008).

The validity of our approach is also easy to understand within this framework: Deleting cell values with measurement error and using MI introduces no biases, and running MI while also using observed cell values that are informative but measured with some error to help inform cell-level priors clearly improves efficency and reduces model dependence. Adopting, instead, an application-specific approach will, under some conditions, perform better, but only with high costs in terms of designing and using specialized statistical models, analyses, and software.[1]

2.2.1 The Foundation: A Multiple Imputation Model

MO builds on MI, which we now review. MI involves using a model to generate multiple imputations for each of the missing cell values (as predicted from all available information in the data set), separate analysis of each of the completed data sets without worry about missing data, and then the application of some easy rules for combining the separate results. The main computational difficulty comes in developing the imputation model.

For expository simplicity, consider a simple special case with only two variables, y_i and x_i ($i = 1, \ldots, n$), where only x_i contains some missing values. These variables are not necessarily outcome and explanatory variables, as they can each play any role in the subsequent analysis model. Everything in this section generalizes to any number of variables and arbitrary patterns of missingness in any or all of the variables (Honaker and King 2010).

We now write down a common model that could be used to apply to the data if they were complete, and then afterwards explain how to use it to impute any missing data scattered through the input variables. This model assumes that the joint distribution of y_i and x_i, $p(y_i, x_i|\mu, \Sigma)$, is multivariate

1. Scholars who have made this connection before have focused almost exclusively on data with validation subsamples, which are relatively rare in the social sciences (Wang and Robins 1998; Brownstone and Valletta, 1996; Cole, Chu, and Greenland, 2006). For a related problem of "editing" data with suspicious cell values, Ghosh-Dastidar and Schafer (2003) develop a MI framework similar in spirit to ours, albeit with an implementation specific to their application.

normal:

$$(x_i, y_i) \sim \mathcal{N}(\mu, \Sigma), \quad \mu = (\mu_y, \mu_x), \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix}, \tag{2.1}$$

where the elements of the mean vector μ and variance matrix Σ are constant over the observations. This model is deceptively simple but powerful: As there is no i subscript on the scalar means μ_x and μ_y, it may appear as though only the marginal means are used to generate imputations. Yet, its joint distribution implies that a prediction is always based on a regression (the conditional expectation) of that one variable on *all* the others, with the population values of the coefficients in the regression a deterministic function of μ and Σ. This is extremely useful in missing data problems for predicting a missing value conditional on observed values. For instance, given model (3.1), the conditional expectation is $E[x_i|y_i] = \gamma_0 + \gamma_1(y_i - \mu_y)$, where $\gamma_0 = \mu_x$ and $\gamma_1 = \sigma_{xy}/\sigma_x$. Researchers have repeatedly demonstrated that this imputation model works as well as more complicated non-linear and non-normal alternatives even for ordinal or categorical variables, and even when more sophisticated models are preferred at the analysis stage (Schafer 1997 and citations in King et al. 2001).

Thus, to estimate the regression of each variable in turn on all the others, we only need to estimate μ and Σ. If there were no missing data, the results would be equivalent to running each of the separate regressions (y_i on x_i and x_i on y_i). But how can we run either of these regressions with arbitrary missing data? The trick, which we now explain, is to find a single set of estimates of μ and Σ from data with scattered missingness, and then to use these to deterministically compute the coefficients of all the separate regressions. To be more specific, the "complete-data" likelihood (i.e., still assuming no missing

data) is simply the product of model (3.1) over the n observations:

$$\mathcal{L}(\theta|y,x) \propto \prod_i p(y_i, x_i|\theta) \tag{2.2}$$

$$= \prod_i p(x_i|y_i,\theta)p(y_i|\theta), \tag{2.3}$$

where $\theta = (\mu, \Sigma)$. (We use variables without an i subscript to denote the vector of observations, so $y = (y_1, \ldots, y_n)$.) This likelihood is not usable as is because it is a function of the missing data, which we do not observe. Thus, we integrate out whatever missing values happen to exist for each observation to produce the actual ("observed-data") likelihood:

$$\mathcal{L}(\theta|y,x^{\text{obs}}) \propto \prod_i \int p(x_i|y_i,\theta)p(y_i|\theta)dx^{\text{mis}} \tag{2.4}$$

$$= \prod_{i\in x^{\text{mis}}} p(y_i|\theta) \prod_{j\in x^{\text{obs}}} p(x_j|y_j,\theta)p(y_j|\theta), \tag{2.5}$$

where x^{obs} denotes the set of cell values in x that are observed and x^{mis} the set that are missing. That we can partition the complete data in this way is justified by the standard "missing at random" (MAR) assumption that the missing values may depend on observed values in the data matrix but not on unobservables (Schafer 1997; Rubin 1976). The key advantage of this expression is that it appropriately assumes that we only see what is actually observed, x^{obs} and y, but can still estimate μ and Σ.[2]

This result enables one to take a large data matrix with scattered missingness across any or all variables and impute missing values based on the regression of each variable on all of the others. The actual imputations used are based on the regression predicted values, their estimation uncertainty (due to the fact that μ and Σ, and thus the calculated coefficients of the regression, are unknown), and the fundamental uncertainty (as represented in the multivariate normal in (3.1) or, equivalently, the

2. This observed-data likelihood is difficult to maximize directly in real data sets with arbitrary patterns of missingness. Fast algorithms to maximize it have been developed that use the relationship between (3.1), (2.4), and the implied regressions, using iterative techniques, including variants of Markov chain Monte Carlo, EM, or EM with bootstrapping.

regression error term from each conditional expectation).

MI works by imputing about five values for each missing cell entry (or more for data sets with unusually high missingness), creating "completed" data sets for each, running whatever analysis model we would have run on the each completed data set as if there were no missing values, and averaging the results using a simple set of rules. The assumption necessary for MI to work properly is that the missing cell values are MAR. This is considerably less restrictive than, for example, the "missing completely at random" assumption required to avoid bias in listwise deletion. See the Appendix for a more formal treatment of the assumptions behind MO.

2.2.2 Incorporating Measurement Error

The measurement error literature uses a variety of assumptions that are, in different ways, more and also less restrictive than our approach. The "classical" error-in-variables model assumes the error is independent of the true value being measured. "Nondifferential" or "surrogate" error is that assumed independent of the dependent variable, conditional on both the true value being measured and any observed pre-treatment predictor variables. Other approaches use assumptions about exclusion restrictions or auxilliary information such as repeated measures. See Imai and Yamamoto (2010) for formal definitions and citations to the literature.

In our alternative approach, we marshall two distinct sources of information to overimpute cell values with measurement error. One could think of cell values with any positive level of measurement error as effectively missing values, and the observed cell value is useless information. In this situation, we can easily translate a measurement error problem into a missing data problem, for which the observed-data likelihood derived in Section 2.2.1 applies directly. The assumption required for this procedure is MAR, which is considerably less restrictive than the assumptions necessary for most prior approaches to dealing with measurement error, and, unlike most other measurement error approaches, it may be used for data sets with arbitrary patterns of measurement error (and missingness) in any

(explanatory or dependent) variables.

Of course, if we think of observations measured with error as reasonable proxies for unobserved values, then treating them as missing will work but may discard valuable information. In fact, variables entirely measured with error may leave no information with which to make (over)imputations under this approach. Thus, we supplement the information that would come from treating cell values measured with error as completely unobserved, and its relatively minimal assumptions, with a second source of information — the proxy measurements themselves along with assumptions about the process by which the proxies are observed. This second source of information enables researchers to make somewhat stronger assumptions, when the measured proxies bear some relationship to the unobserved true values, in return for considerably more efficient estimates.

For expository clarity, we continue, without loss of generality, our simple two-variable example from the previous section. Thus, let y_i be a single fully observed cell value and x_i^* be a true but unobserved cell value (these variables may serve any role in a subseqent analysis model), with $(y_i, x_i^*) \sim \mathcal{N}(\mu, \Sigma)$ as above. To this, we add an observed w_i which is a proxy, measured with error, for x_i^*. For expository simplicity, we focus on the case with no (fully) missing values, which in this context would be unobserved cell values without corresponding proxy values.

With this setup, we describe the second source of information in our approach as coming from the specification of a specific probability density to represent the data generation process for the proxy w_i. This, of course, is an assumption and we allow a wide range of choices, subject to two conditions, one technical and one substantive. First, the class of allowable data generation processes in our approach involves any probability density that possesses the property of *statistical duality*. This is a simple property (related to self-conjugacy in Bayesian analysis) possessed by a variety of distributions, such as normal, Laplace, Gamma, Inverse Gamma, Pareto, and others (Bityukov et al. 2006).[3] (We use this

3. If a function $f(a, b)$ can be expressed as a family of probability densities for variable a given parameter b, $p(a|b)$, and a family of densities for variable b given parameter a, $p(b|a)$, so that $f(a, b) = p(a|b) = p(b|a)$, then $p(a|b)$ and $p(b|a)$ are said to be statistically dual.

property to ease implemention in Section 2.2.3.) Second, we require that the mean (or an additive function of the mean) of the distribution be the unobserved true cell value x_i^*, and that the parameters of the distribution are distinct from the complete-data parameters, θ, and are known or separately estimated.

A simple special case of this data generation process is random normal measurement error, $w_i \sim \mathcal{N}(x_i^*, \sigma_u^2)$, with σ_u^2 set to a chosen or estimated value (we discuss interpretation and estimation of σ_u^2 in Section 2.3). Other special cases allow for heteroskedastic measurement error, such as might occur with GDP from a country where a government's statistical office is professionalizing over time; mortality statistics from countries with and without death registration systems; or survey responses from a self-report vs elicited about that person from someone else in the same household. This approach can handle biased measurement error, where $\mathrm{E}[w_i|x_i^*] = a_i + x_i^*$, so long as the bias, a_i, is known or estimable. For instance, if validation data is available, a researcher could estimate the bias of the measure or use a model to estimate how the offset changes with observed variables. From our perspective, a cell value (or variable) that doesn't possess at least this minimally known set of relationships to its true value could more easily be considered a new observation of a different variable rather than a proxy for an unobserved one.[4] When the bias is not known and cannot be estimated, we are left with a class of data generation processes (rather than a single one) for the proxy; this results under our procedure in a "robust Bayesian" *class* of posteriors (rather than a single Bayesian posterior), from which overimputations may be drawn (Berger 1994; King and Zeng 2002).

The result of these assumptions is a complete-data likelihood that can be used to encompass both

4. If the relationship between the underlying variable and its mismeasurement is completely unknown, a different approach may be required. For instance, structural equation modeling or factor analysis is sometimes appropriate if a large set of measures all capture some aspect of an unobserved concept.

methodological problems:

$$L(\theta, \sigma_u^2 | y, w, x^*) \propto \prod_i p(y_i, w_i, x_i^* | \theta, \sigma_u^2) \tag{2.6}$$

$$= \prod_i p(w_i | x_i^*, y_i, \theta, \sigma_u^2) p(x_i^* | y_i, \theta) p(y_i | \theta) \tag{2.7}$$

$$= \prod_i p(w_i | x_i^*, y_i, \sigma_u^2) p(x_i^* | y_i, \theta) p(y_i | \theta). \tag{2.8}$$

The first equality uses the rules of conditional probability; the key assumption (needed for the second equality) is expressed here by the density for w_i not depending on the parameters of the overall likelihood, $\theta = (\mu, \Sigma)$: $p(w_i | x_i^*, y_i, \theta, \sigma_u^2) = p(w_i | x_i^*, y_i, \sigma_u^2)$. Note that (2.8) is identical to the complete-data likelihood in MI model (2.3), with the additional factor, $p(w_i | x_i^*, y_i, \sigma_u^2)$, for the proxy's data generation process. (To generate the observed-data likelihood in this case would of course require the analogous integration as in (2.4), which we omit here to save space. See Appendix A for a full description.)

We may sometimes wish to further simplify and assume normal error, $p(w_i | x_i^*, y_i, \theta, \sigma_u^2) = \mathcal{N}(x_i^*, \sigma_u^2)$, with a chosen or estimated value of the variance of the measurement error σ_u^2. When σ_u^2 is small we have a reasonable precision in our estimate of the location of x_i^*. As the size of the measurement error grows, w_i reveals less information about the true value of x_i^*. Heuristically, as σ_u^2 becomes infinite, w_i tells us nothing, and we may as well discard it from the data set and treat it as missing. In this limiting case, where no information is directly observed about x_i^*, then $\lim_{\sigma_u^2 \to \infty} p(w_i | x_i, \sigma_u^2)$ approaches a constant and the complete-data likelihood (2.8) becomes proportional to the model for missing data alone (2.3). This proves that the most commonly used model for missing data is a limiting special case of our approach.

2.2.3 Implementation

In a project designed for an unrelated purpose, Honaker and King (2010) propose a fast and computationally robust MI algorithm that allows for informative Bayesian priors on missing individual cell values. The algorithm is known as EMB, or EM with bootstrapping. They use this model to incorporate qualitative case-specific information about missing cells to improve imputations. To make it easy to implement our approach, we prove in Appendix A that the same algorithm can be used to estimate our model. The statistical duality property assumed there enables us to turn the data generation process for w_i into a prior on the unobserved value x_i^*, without changing the mathematical form of the density. For example, in the simple random normal error case, the data generation process for w_i is $\mathcal{N}(x_i^*, \sigma_u^2)$ but, using the property of statistical duality of the normal, this is equivalent to a prior density for the unobserved x_i^*, $\mathcal{N}(w_i, \sigma_u^2)$. This result shows that we can use the existing EMB algorithm.

This strategy also offers important intuitions: our approach can be interpreted as treating the proxy variables as informative, observation-level means (or functions of the means) in priors on the unobserved missing cell values. Our imputations of the missing values, then, will be precision-weighted combinations of the proxy variable and the predicted value from the conditional expectation (the regression of each variable on all others) using the missing data model. In addition, the parameters of this conditional expectation (computed from μ and Σ) are informed and updated by the priors on the individual cell values.

Under our overall approach, then, all cells in the data matrix with measurement error are replaced — overwritten in the data set, or *overimputed* in our terminology — with multiple overimputations that reflect our best guess and uncertainty in the location of the latent values of interest x_i^*. These overimputations include the information from our measurement error model, or equivalently the prior with mean set to the observed proxy variable measured with error, as well as all predictive information available in the observed variables in the data matrix. At the same time, all missing values are imputed. The same procedure is used to fill in multiple completed data sets; usually about five data sets is

sufficient, but more may be necessary with large fractions of missing cells or high degrees of measurement error. Imputations and overimputations vary across the multiple completed data sets — with more variation when the predictive ability of the model is smaller and measurement error is greater — while correctly observed cell values remain constant.

Researchers create a collection of completed data sets once and then run as many analyses of these as desired. The same analysis model is applied to each of the completed (imputed and overimputed) data sets as if it were fully observed. A key point is that the analysis model need not be linear-normal even though the model for missing values and measurement error overimputation is (Meng 1994). The researcher then applies the usual MI rules for combining these results (see Appendix A).

2.2.4 Monte Carlo Evidence

We now offer Monte Carlo evidence for MO, using a data generation process that would be difficult or impossible for most prior approaches to measurement error. We use two mismeasured variables, a non-normal dependent variable, scattered (but not completely random) missing data, and a nonlinear analysis model. The measurement error is independent random normal with variances that each account for 25% of the total variance for each proxy, meaning these are reasonably noisy measures. In doing so, we attempt to recreate a difficult but realistic political science data situation, with the addition of the true values so we can use them to validate the procedure.

In a real application, a researcher may only have a rough sense of the measurement error variances. We thus run our simulations assuming a range of levels for these variances, holding their true value fixed, to see how these differing assumptions affect estimation. (In the next section, we discuss how to interpret or estimate this variance.)

We generated proxies x and z for the true variables x^* and z^*, respectively, using a normal data generation process with the true variables as the mean and a variance equal to $\sigma_u^2 = \sigma_v^2 = 0.5$.[5] At each

5. We let y_i, the dependent variable of the analysis model, follow a Bernoulli distribution with probability $\pi_i = 1/(1 + \exp(-X_i\beta))$, where $X_i = (x_i^*, z_i^*, s_i)'$ and $\beta = (-7, 1, 1, -1)$. We allow scattered missingness of a random 10% of the all cell

combination of σ_u^2 and σ_v^2, we calculate the mean square error (MSE) for the logit coefficients of the overimputed latent variables. We took the average MSE across these coefficients and present the results in Figure 2.1. On the left is the MSE surface with the error variances on the axes along the floor and MSE on vertical axis; the right graph shows the same information viewed from the top as a contour plot.

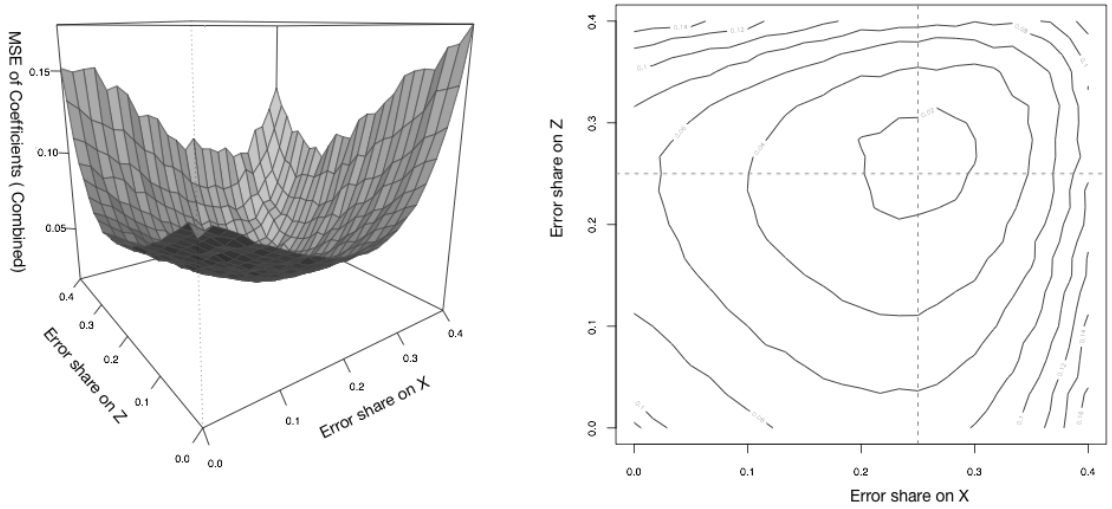


Figure 2.1: On the left is a perspective plot of the mean square error of a logit analysis model estimates after multiple overimputation with various assumptions about the measurement error variance. The right shows the same information as a contour plot. Note that the axes here are the share of the observed variance due to measurement error which has a true value of 0.25, which is precisely where the MSE reaches a minimum.

The figure shows that when we assume the absence of measurement error (i.e., $\sigma_u^2 = \sigma_v^2 = 0$), as most researchers do, we are left with high MSE values. As the assumed amount of measurement error grows, we see that the MO lowers the MSE smoothly. The MSE reaches a minimum at the true value of the measurement error variance (the gray dotted lines in the contour plot).[6] Assuming values that are much

values of y, x, and z when (the fully observed) s is greater than its mean. We created the true, latent data (x^*, z^*, s) by drawing from a multivariate normal with mean vector $(5, 3, 1)$ and covariance matrix $(1.5\ 0.5\ -0.2,\ 0.5\ 1.5\ -0.2,\ -0.2\ -0.2\ 0.5)$.

6. In this simulation, the variance of the estimates is swamped by the squared bias of the estimates, so that any difference in the MSE is almost entirely due to bias, rather than efficiency. More succinctly, these plots are substantively similar if we replace MSE with bias.

too high also leads to larger MSEs, but the figure reveals one of the types of robustness of the MO procedure in that a large region exists where MSE is reduced relative to the naive model assuming no error, and so one need not know the measurement error variance except very generally. We discuss this issue further below.

CATEGORICAL VARIABLES MEASURED WITH ERROR While our imputation model assumes the data is drawn from a multivariate normal distribution, non-normal variables, such as categorical variables, can be included in the imputation and can even be overimputed for measurement error. It is well known in the multiple imputation literature that imputation via a normal model works well for categorical variables, and indeed as well as models designed especially for categorical variables and even when the analysis model is nonlinear (Schafer 1997; Schafer and Olsen 1998). Our own detailed simulations (not presented here) confirm that this standard result for MI also applies to MO.

2.2.5 COMPARISON TO OTHER MEASUREMENT ERROR CORRECTION TECHNIQUES

Measurement error is a core threat to statistical analysis and many have proposed solutions to its problems. These solutions broadly fall into two camps: general-purpose methods and application-specific methods. General-purpose methods are easily implemented across a wide variety of models, while application-specific methods are closely tailored to a particular context. For more information about the various approaches to measurement error, see Fuller (1987) and Carroll, Ruppert, and Stefanski (1995).

The first general-purpose method, *regression calibration* (Carroll and Stefanski 1990), is similar in spirit to MO in that it replaces the observed, mismeasured variable with an estimate of the underlying unobserved variable and then performs the desired analysis on this "calibrated data." In fact, one can think of MO as an combination of regression calibration and multiple imputation, two methods previously thought as distinct and in competition with one another (Cole, Chu, and Greenland, 2006). As White (2006) points out, multiple imputation relies on validation data and ignores any replicate

measures and regression calibration ignores any validation data completely. MO combines the best parts of each of these approaches by utilizing all information when it is available.

The easiest technique to implement is a simple method-of-moments estimator, which exploits the exact relationship between the biased estimates and the amount of measurement error present in the data. This estimator simply corrects a biased estimate of a linear regression coefficient by dividing it by the reliability ratio, $\sigma^2_{x^*}/\sigma^2_w$. This technique depends heavily on the estimate of the measurement error variance and, in our simulations, has poor properties when this estimate is incorrect. Further, the method-of-moment technique requires the analysis model to be linear.

Other general approaches to measurement error include simulation-extrapolation, or SIMEX, (Cook and Stefanski 1994), and minimal-assumption bounds (Leamer 1978; Klepper and Leamer 1984; Black, Berger, and Scott 2000). These are both excellent approaches to measurement error, but they both have features that limit their general applicability. SIMEX is designed to simulate the effect of adding *additional* measurement error to a single mismeasured variable, then use these simulations to extrapolate back to the case with no measurement error. In situations with multiple mismeasured variables, SIMEX becomes harder to compute and more dependent on the extrapolation model. The minimal-assumption bounds are useful for specifying a range of parameter values consistent with a certain set of assumptions on the error model. Bounds typically require fewer assumptions than our multiple overimputation model, but obviously eliminate the possibility of point estimation. A comprehensive approach to measurement error could utilize minimal-assumption bounds with the overimputation bounds and point estimates below.

Structural equation modeling (SEM) attempts to solve the problem of measurement error in a different way.[7] The goal of SEM is to find latent dimensions that could have generated a host of observed measures, while our goal is to rid a particular variable (or variables) of its measurement error. While discovering and measuring latent concepts is a useful and common task in political science, there are

7. Lee (2007) covers a number of Bayesian approaches to structural equation modeling, including some that take into consideration missing data.

many cases in which we want to measure the effect of a specific variable and measurement error stands in the way. SEM would sweep that variable up into a larger construct and perhaps muddle the question at hand. Thus, MO is not so much a replacement for SEM, but rather an approach to a different set of substantive questions. Furthermore, MO is better equipped to handle gold-standard and validation data since it is unclear how to incorporate these into a structural equation modeling framework.

2.3 Specifying or Estimating the Measurement Error Variance

The measurement error variance is unidentified in our approach and all others, without some further data or assumptions (Stefanski, 2000). When little or no extra information is available, we show how to reparametrize σ_u^2 to a scale that is easier to understand and how we can provide bounds on the quantity of interest (Section 2.3.1). When replicated correlated proxies are available, we show how to estimate σ_u^2 directly (Section 2.3.2). And finally we show how to proceed when σ_u^2 varies over the data set or when gold standard observations are available (Section 2.3.3).

2.3.1 Interpretation through Reparametrization and Bounding

Section 2.2.4 shows that using the true measurement error variance σ_u^2 with MO will greatly reduce the bias and MSE relative to the usual procedure of making believe measurement error does not exist (which we refer to as the "denial" estimator). Moreover, in the simulation presented there (and in others we have run), the researcher needs only have a general sense of the value of these variances to greatly decrease the bias of the esimates. Of course, knowing the value of σ_u^2 (or σ_u) is not always obvious, especially on its given scale. In this section, we deal with this problem by reparameterizing it into a more intuitive quantity and then putting bounds on the ultimate quantity of interest.

The alternative parametrization we have found useful is the *proportion of the proxy variable's observed variance due to measurement error*, which we denote $\rho = \frac{\sigma_u^2}{\sigma_x^2+\sigma_u^2} = \frac{\sigma_u^2}{\sigma_w^2}$, where σ_w^2, the variance of our proxy. This is easy to calculate directly if the proxy is observed for an entire variable (or at least more

than one cell value). Thus, if we know the extent of the measurement error, we can create an estimated version of $\hat{\sigma}_u^2 = \rho\hat{\sigma}_w^2$ and substitute it for σ_u^2 in the complete-data likelihood (2.8).

In Figure 2.2, we present Monte Carlo simulations of how our method works when we alter our assumptions on the scale of ρ rather than σ_u^2.[8] More importantly, it shows how providing little or no information about the measurement error can bound the quantities of interest. Leamer (1978, pp. 238–243) showed that we can use a series of reverse regressions in order to bound the true coefficient without making any assumptions about the amount of measurement error. We compare these "minimal-assumption" bounds to the more model-based multiple overimputation bounds. The vertical axis in the left panel is the value of the coefficient of a regression of the overimputed w on y. The orange points and vertical lines are the estimates and 95% confidence intervals from overimputation as we change our assumption about ρ on the horizontal axis.

We can see that the denial estimator, which treats w as if it were perfectly measured (in red), severely underestimates the effect calculated from the complete data (solid blue horizontal line), as we might expect from the standard attenuation result. As we assume higher levels of ρ with MO, our estimates move smoothly toward the correct inference, hitting it right when ρ reaches its true value (denoted by the vertical dashed line). Increasing ρ after this point leads to overcorrections, but one needs to have a very bad estimate of ρ to make things worse than the denial estimator. The root mean square error leads to a similar conclusion and is thus also minimized at the correct value of ρ.

A crucial feature of MO is that it can be informative even if one has highly limited knowledge of the degree of measurement error. To illustrate this, the left panel of Figure 2.2 offers two sets of bounds on the quantity of interest, each based on different assumptions about ρ. We use the reverse regression technique of Leamer (1978) to generate minimal-assumption bounds, which make no assumptions about ρ (the mean of these bounds are in light tan). In practice, it would be hard to justify using a

8. For these simulations, we have $y_i = \beta x_i + \varepsilon_i$ with $\beta = 1$, $\varepsilon_i \sim \mathcal{N}(0, 1.5^2)$, $x_i^* \sim \mathcal{N}(5, 1)$, and $\sigma_u^2 = 1$. Thus, we have $\rho = 0.5$. We used sample sizes of 1,000 and 10,000 simulations

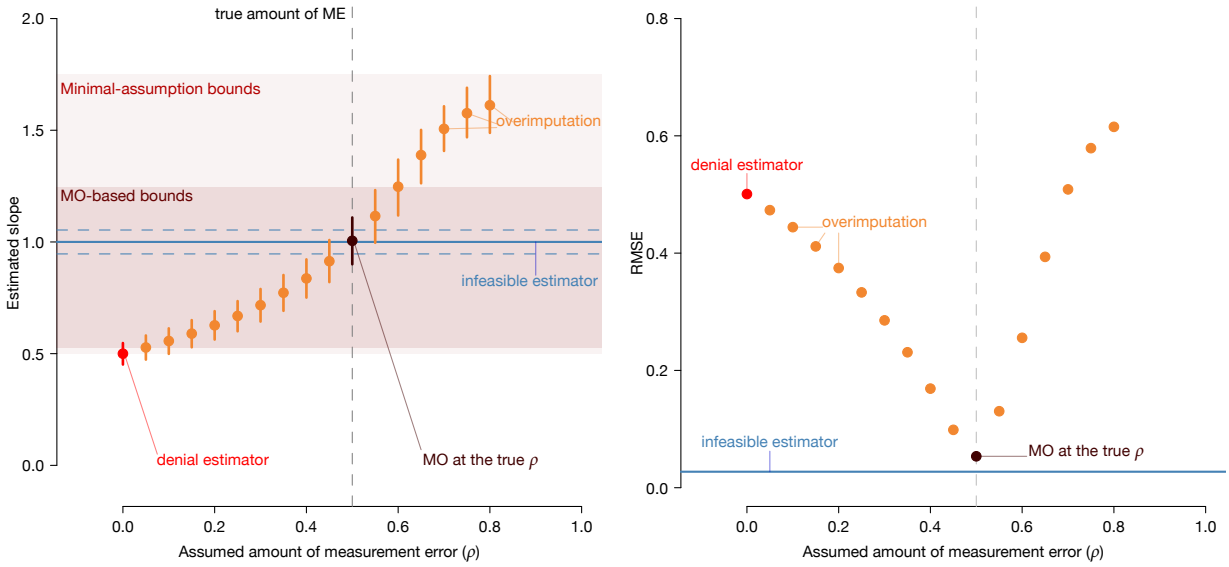


Figure 2.2: Simulation results using the denial estimator (that assumes no measurement error, in red), the complete-data, infeasible estimator (in blue), and the MO estimator (in orange), with varying assumptions about the degree of mismeasurement. The MO estimator at the correct value of ρ is in dark red. The left panel shows estimates of the coefficients of interest along with confidence bands. In the background, the light tan area shows the minimal-assumption bounds and the dark tan region gives bounds assuming $\rho \in [0.05, 0.6]$. The right panel shows MSE for the same range of estimates.

variable with over 80% measurement error, but even in this extreme situation the bounds on the quantity of interest do convey a great deal of information. They indicate, for example, that the denial estimator is an underestimate of the quantity of interest and almost surely within approximately the range [0.5,1.75]. Note that all of our MO estimates are within these bounds. In simulation in which we lowered the true ρ, we have found that even dramatic overestimates of ρ still lead to MO estimates that obey these bounds.[9]

Alternatively, we might consider making a more informative (and reasonable) assumption about ρ. Suppose that we know that there is some positive measurement error, but that less than 70% of the observed variance is due to measurement error. These are informative assumptions about ρ and allow MO to estimate bounds on the estimated coefficient. The result is that the bounds shrink (in dark tan, marked "MO-based") closer around the truth. MO thus helps researchers learn about how various assumptions about measurement error affect their estimates.[10] These bounds depend on the imputation model, but because MO allows for arbitrary patterns of mismeasurement, this bounding procedure is extraordinarily flexible. The MO-based bounding approach to measurement error shifts the burden from choosing the correct share of measurement error to choosing a range of plausible shares. Researchers may feel comfortable assuming away higher values of ρ since we may legitimately consider a variable with, say, 80% measurement error as a different variable entirely. The lower bound on ρ can often be close to 0 in order to allow for small amounts of measurement error. [11]

This figure also highlights the dangers of incorrectly specifying ρ. As we assume that more of the proxy is measurement error, we eventually overshoot the true coefficient and begin to see increased MSE. Note, though, that there is again considerable robustness to incorrectly specifying the prior in this

9. More generally, simulations run at various values of the true ρ lead to the same qualitative results as presented here. Underestimates of ρ lead to underestimates of the true slope and overestimate of ρ lead to overestimates of the true slope.

10. If we use MO at all levels of ρ to generate the most assumption-free MO-based bounds possible, the bounds largely agree with the minimal-assumptions bounds.

11. These simulations also point to a use of MO as tool for sensitivity analysis. MO not only provides bounds on the quantities of interest, but can provide what the estimated quantity of interest would be under various assumptions about the amount of measurement error.

case. Any positive value ρ does better than the naive estimator until we assume that almost 70% of the proxy variance is due to error. This result will vary, of course, with the true degree of measurement error and the model under study.

2.3.2 Estimation with Multiple Proxies

When multiple proxies (or "repeated measures") of the same true variable are available, we can use relationships among them to provide point estimates of the required variances, and to set the priors in MO. For example, suppose for the same true variable x^* we have two unbiased proxies with normal errors that are independent after conditioning on x^*:

$$w_1 = x^* + u: \; u \sim N\big(0, \sigma_u^2\big), \qquad w_2 = ax^* + b + v: \; v \sim N\big(0, (c\,\sigma_u)^2\big) \tag{2.9}$$

where a, b, c are unknown parameters, that rescale the additional proxy measure to a different range, mean, and different degree of measurement error. The covariances and correlations between these proxies can be solved as $\mathrm{E}\big[\mathrm{cov}(w_1, w_2)\big] = a\,\mathrm{var}(x^*)$ and $\mathrm{E}\big[\mathrm{cor}(w_1, w_2)\big] = \gamma\,\mathrm{var}(x^*)/\mathrm{var}(w_1)$, where a is one of the scale parameters above, and γ is a ratio:

$$\gamma^2 = a^2\frac{\mathrm{var}(w_1)}{\mathrm{var}(w_2)} = \frac{\mathrm{var}(x^*) + \mathrm{var}(u)}{\mathrm{var}(x^*) + (c^2/a^2)\mathrm{var}(u)} \tag{2.10}$$

If the measurement error is uncorrelated with x^* the variances decompose as $\sigma_u^2 = \sigma_{w_1}^2 - \sigma_{x^*}^2$. This leads to two feasible estimates of the error variances for setting priors. First:

$$s^2(u) = \mathrm{var}(w_1) - \mathrm{cov}(w_1, w_2) = \mathrm{var}(w_1) - \mathrm{var}(x^*)\,a \tag{2.11}$$

which is exactly correct when $a = 1$, that is, when w_2 is on the same scale (with possibly differing intercept) as w_1. Similarly,

$$s^2(u) = \text{var}(w_1)\big(1 - \text{cor}(w_1, w_2)\big) = \text{var}(w_1) - \text{var}(x^*)\ \gamma \tag{2.12}$$

which is exactly correct when $c = a \Leftrightarrow \gamma = 1$, that is, the second proxy has the same relative proportion of error as the original proxy.

2.3.3 Estimation with Heteroskedastic Measurement Error

In some applications, the amount of measurement error may vary across observations. Although most corrections in the literature ignore this possibility, it is easy to include in the MO framework, and doing so often makes estimation easier. To include this information, merely add a subscript i to the variance of the measurement error: $p(w_i|x_i^*, \sigma_{ui}^2) = \mathcal{N}(w_i|x_i^*, \sigma_{ui}^2)$. We consider two examples.

First, suppose the data include some observations measured with error and some without error. That is, for fully observed data points, let

$w_i = x_i^*$, or equivalently $\sigma_{ui}^2 = 0$. This implies that $p(w_i|x_i^*)$ drops out of the complete-data likelihood and x_i^* becomes an observed cell. Then the imputation model would only overimpute cell values measured with error and leave the "gold-standard" observations as is. If the other observations have a common error variance, σ_u^2, then we can easily estimate this quantity, since the variance of the gold-standard observations is σ_x^2 and the mismeasured observations have variance $\sigma_x^2 + \sigma_u^2$. This leads to the feasible estimator,

$$\hat{\sigma}_u^2 = \hat{\sigma}_{mm}^2 - \hat{\sigma}_{gs}^2, \tag{2.13}$$

where $\hat{\sigma}_{mm}^2$ is the estimated variance of the mismeasured observations and $\hat{\sigma}_{gs}^2$ is the estimated variance of the gold-standard observations.[12]

12. This logic assumes that the gold-standard observations are a random sample of the observations. When this assumption is implausible, we can use the reparameterization approach of Section 2.3.1.

As second special case of heteroskedastic measurement error, MO can handle situations where the variance is a linear function of another variable. That is, when $\sigma_{ui}^2 = rZ_i$, where Z_i is variable and r is the proportional constant relating the variable to the error variance. If we know r (or we can estimate it through variance function approaches), then we can easily incorporate this into the prior above using $p(w_i|x_i^*, r, Z_i) \sim \mathcal{N}(w_i|x_i^*, rZ_i)$.

2.4 Correlated Proxies

In this section and the next, we show how MO is robust to data problems that may occur in a large number of settings and applications. We show here how MO is robust to theoretical measurement dilemmas that occur regularly in political science data. In the sequel, we show more pragmatic robustness to a number of measurement applications in real data.

Until now we have assumed that measurement error is independent of all other variables. We now show how to relax this assumption. Many common techniques for treating measurement error make this strong assumption and are not robust when it is violated. For example, probably the most commonly implemented measurement error model (in the rare cases that a correction is attempted at all) is the classic errors-in-variables (EIV) model. We thus first briefly describe the EIV model to illustrate the strong assumptions required. The EIV model is also a natural point of comparison to MO, since both can be thought of as replacing mismeasured observations with predictions from auxiliary models.

2.4.1 The Foundation: The Errors-in-variables Model

As before, assume y_i and x_i^* are jointly normal with parameters as in (3.1). Suppose instead of x^* we have a set of proxy variables which are measures of x^* with some additional normally distributed

random noise:

$$w_{i1} = x_i^* + u_i, \qquad u_i \sim \mathcal{N}(0, \sigma_u^2); \tag{2.14}$$

$$w_{i2} = x_i^* + v_i, \qquad v_i \sim \mathcal{N}(0, \sigma_v^2); \tag{2.15}$$

ordered such that $\sigma_u^2 < \sigma_v^2$, making w_1 the superior of the two proxies as it has less noise.

Suppose the true relationship is $y_i = \alpha x_i^* + \varepsilon_{i1}$, and we instead use the best available proxy and estimate $y_i = \beta w_{i1} + \varepsilon_{i2} = \beta(x_i^*) + \beta(u_i) + \varepsilon_{i2}$. We then get some degree of attenuation $0 < \beta < \alpha$ since the coefficient on u_i should be zero. This attenuation is shown in one example in the right of Figure 2.3 where the relationship between y and w_1 shown in red is weaker than the true relationship with x^* estimated in the left graph and copied in black on the right.

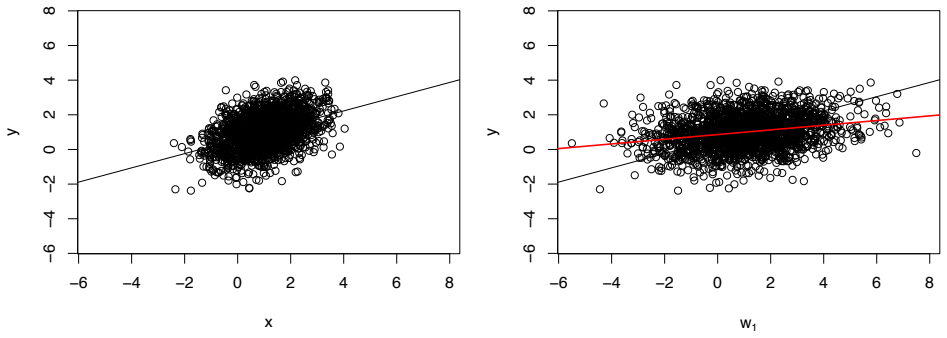


Figure 2.3: On the left we see the true relationship between y and the latent x^*. When the mismeasured proxy w_1 is used instead, the estimated relationship (shown in red) is attenuated compared to the true relationship (shown in black in both graphs).

In this simple example we can calculate the expectation of this attenuation. The coefficient on w_{i1} will be

$$\mathrm{E}[\hat{\beta}_1] = \mathrm{E}\left[\frac{\sum_i (x_i^* + u_i - (\overline{x^* + u}))(y_i - \bar{y})}{\sum_i (x_i^* + u_i - (\overline{x^* + u}))^2}\right] = \frac{\sum_i (x_i^* - \bar{x}^*)(y_i - \bar{y})}{\sum_i (x_i^* - \bar{x}^*)^2 + \sigma_u^2}, \tag{2.16}$$

where $\overline{x^* + u}$ and $\bar{x}^*$ are the sample means of w_1 and x^*, respectively. The last term in the denominator,

σ_u^2, causes this attenuation. If the variance of the measurement error is zero the term drops out and we get the correct estimate. As the measurement error increases, the ratio tends to zero.

The coefficients in the EIV approach can be estimated either directly or in two stages. A two-stage estimation procedure is the common framework to build intuition about the model and the role of the additional proxy measure. In this approach, we first obtain estimates of x^* from the relationship between the w's since they only share x^* in common, $\hat{w}_{i1} = \hat{\gamma} w_{i2}$, and then use these predictions to estimate $y_i = \delta \hat{w}_{i1} + \varepsilon_{i3}$, where now $\hat{\delta}$ is an unbiased estimate of α. The relationship between the two proxy variables is shown in the left of Figure 2.4, and the relationship between the first stage predicted values of w_1 and y is shown in green in the right figure. This coincides almost exactly with the true relationship still shown in black in this figure.

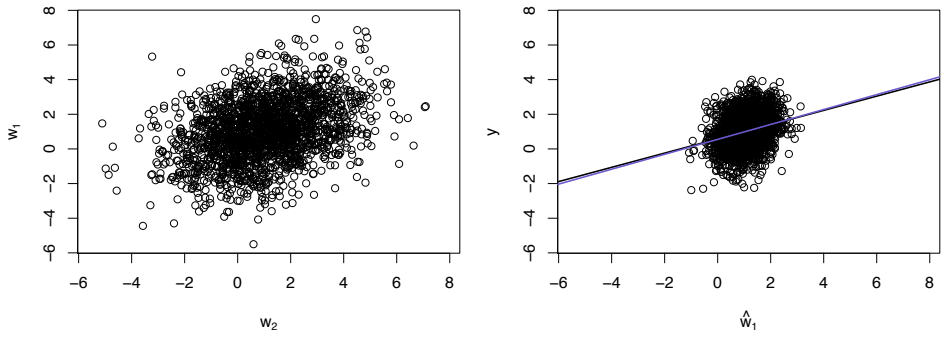


Figure 2.4: The relationship between two mismeasured proxy variables (left), and the relationship between the predicted values from this model and y (right). The relationship here, shown in green, recovers the true relationship, shown in black.

In Figure 2.5 we illustrate how the EIV model performs in data that meet its assumptions. The black distributions represent the distribution of coefficients estimated when the latent data x^* is available in a simulated data set of size 200.[13] The naive regressions that do not account for measurement error are shown in red in both graphs. The coefficient on w_1 is attenuated to towards zero (bottom panel). The estimated constant term is biased upwards to compensate (top panel). In each simulated data set, we

13. In these simulations, $n = 200$, $(x^*, y) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1)$, $\Sigma = (1\ 0.4, 0.4\ 1)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$.

use the EIV model (in green), and see that the distribution of estimated parameters using the proxies resembles the distribution using the latent data, although with slightly greater variance. Thus there is some small efficiency loss, but the EIV model clearly recovers unbiased estimates when its assumptions are met.

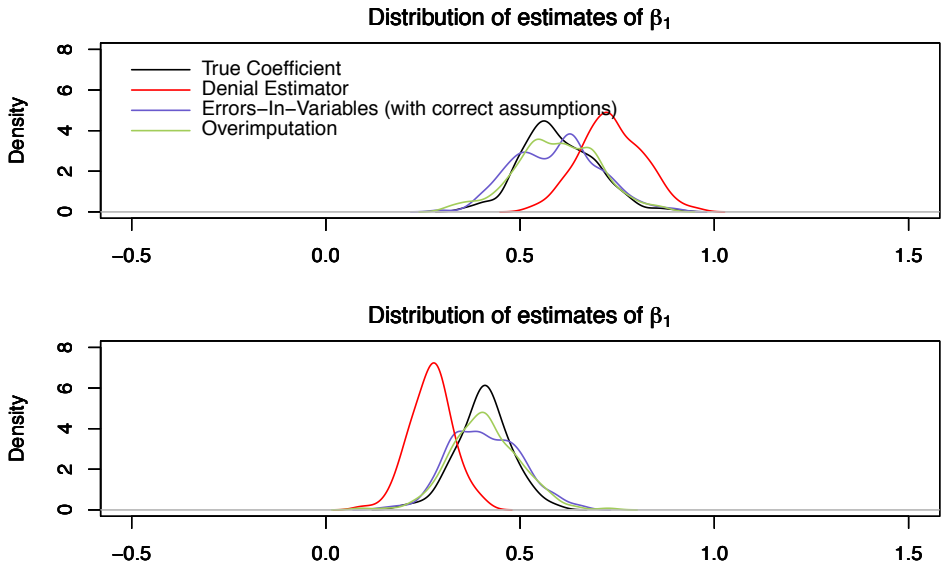


Figure 2.5: Coefficients estimated from variables with measurement error (shown in red) attenuate the effect of the independent variable towards zero, and also bias the constant in compensation. The estimates recovered from the EIV model (in green) recover the true distribution, but are of course less efficient (slightly higher variance) than the original latent data (in black).

We also run the MO on the same simulated data sets in which we ran the EIV model. The distribution of coefficients (which we present below) recovers the distribution that would have been estimated if the latent data had been available. Thus, in the simple setting where the assumptions of the EIV model are met, our approach performs equivalently.

2.4.2 Robustness to Violating Assumptions

If we think of the coefficient on x^* as the ratio of $\text{cov}(x^*, y)$ to $\text{var}(x^*)$, then the attenuation in equation (2.16) is being driven by the fact that $\text{var}(w_1) > \text{var}(x^*)$ because of the added measurement error. Therefore $\text{var}(w_1)$ is not a good estimate of $\text{var}(x^*)$, even though $\text{cov}(w_1, y)$ *is* a good measure of $\text{cov}(x^*, y)$. With this in mind, the numerically simpler—but equivalent—one stage approach to the errors-in-variables model has a useful intuition. We substitute $\text{cov}(w_1, w_2)$ as an estimate of $\text{var}(x^*)$ because w_1, w_2 only covary through x^*. Thus we have as our estimate of the relationship:[14]

$$\hat{\delta} = \frac{\sum_i (w_{i1} - \bar{w}_1)(y_i - \bar{y})}{\sum_i (w_{i1} - \bar{w}_1)(w_{i2} - \bar{w}_2)} = \frac{\sum_i (x_i^* - \bar{x}^*)(y_i - \bar{y}) + u_i(y_i - \bar{y})}{\sum_i (x_i^* - \bar{x}^*)^2 + u_i(x_i^* - \bar{x}^*) + v_i(x_i^* - \bar{x}^*) + u_i v_i}. \tag{2.17}$$

In order to recover the true relationship between x^* and y we need the last term in the numerator and the last three in the denominator to drop out of equation (2.17). To obtain a consistent estimate, then, EIV requires: (1) $\text{E}(u_i \cdot y_i) = 0$, (2) $\text{E}(u_i \cdot x_i^*) = 0$ and $\text{E}(v_i \cdot x_i^*) = 0$, and (3) $\text{E}(u_i \cdot v_i) = 0$. Indeed, when these conditions are not met the resulting bias in the EIV correction can easily be larger than the original bias caused by measurement error. However, as we now show in the following three subsections, MO is robust to violations of all but the last condition.

Measurement error correlated with y

The first of the conditions for EIV to work is that the measurement error is unrelated to the observed dependent variable. As an example of this problem, we might think that infant mortality is related to international aid because donors want to reduce child deaths. If countries receiving aid are intentionally underreporting infant mortality, to try to convince donors the aid is working, then the measurement error in infant mortality is negatively correlated with the dependent variable, foreign aid. If instead countries searching for aid are intentionally overreporting infant mortality as a stimulus for receiving

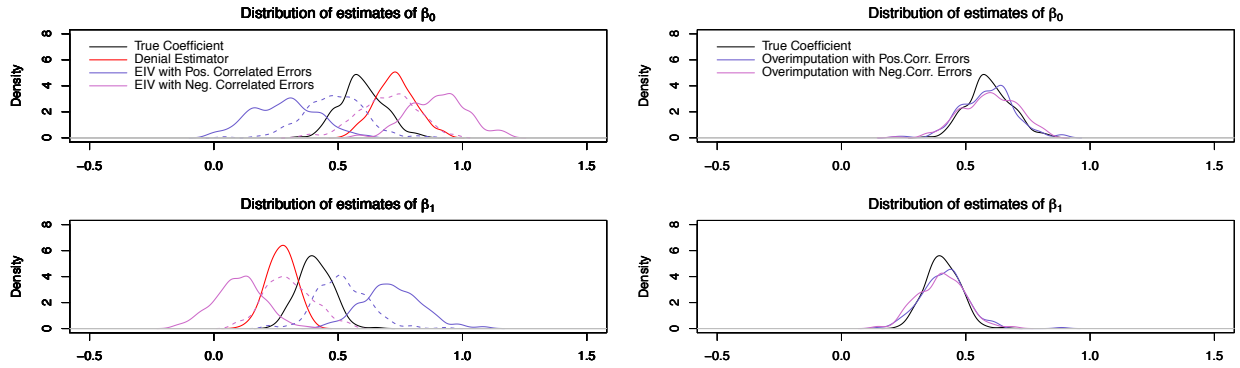
14. In a multivariate setting this becomes $\hat{\delta} = (W_1' W_2)^{-1} W_1' Y$ where W_j is the set of regressors using the j-th proxy measure for x^*.

aid, then measurement error is positively correlated with the dependent variable. Both scenarios are conceivable. This problem with the errors-in-variables approach is well known, because the errors-in-variables model has an instrumental variables framework, and this is equivalent to the problem of the instrument being exogenous of y in the more common usage of instrumental variables as a treatment for endogeneity.

In Figure 2.6a we demonstrate this bias with simulated data.[15] The violet densities show the distribution of parameter estimates when there is negative correlation of 0.1 (dashed) and 0.3 (solid) between the measurement error and the dependent variable. In the latter case the bias in the correction has exceeded the original bias from measurement error, still depicted in red. The blue densities show that positive correlation of the errors create bias of similar magnitude in the opposite direction. Again, the size of the bias can be greater than that originally produced by the measurement error we were attempting to correct. Moreover, the common belief with measurement issues is that any resulting bias attenuates the coefficients so that estimates are at least conservative, however, here we see that the bias in the error-in-variables approach can actually exaggerate the magnitude of the effect.

We now analyze the same simulated data sets with MO. To apply the MO model, we estimate the measurement error variance from the correlation between the two proxies and leave the mean set to the better proxy. As Figure 2.6b indicates, MO recovers the distribution of coefficients for each of the data generation processes: The green line represents the distribution when there is no correlation. The violet line represents the distribution when there is positive correlation. The blue line (barely visible under the other two) represents the distribution with negative correlation. All three distributions are close to each other and close to the true distribution in black using the latent data.

15. In these simulations, similar to previous, $n = 200$, $(x^*, y, u, v) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 0, 0)$, $\Sigma = (1\ 0.4\ 0\ 0, 0.4\ 1\ 0\ \rho, 0\ 0\ \sigma_u^2\ 0, 0\ \rho\ 0\ \sigma_v^2)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$. Thus, the measurement errors are drawn at the same time as x^* and y with mean zero. While ρ allows the error, v, to covary with y, and across the simulations it is set as one of $\rho \in \{-0.3, -0.1, 0.1, 0.3\}$. The observed mismeasured variables are constructed as $w_1 = x^* + u$, $w_2 = x^* + v$.



(a) Estimates from Errors-in-Variables **(b)** Estimates from Multiple Overimputation

Figure 2.6: With data generated so that proxy variables are correlated with the dependent variable, EIV (left graphs) gives biased estimates whereas MO (right graphs) gives robust, unbiased estimates.

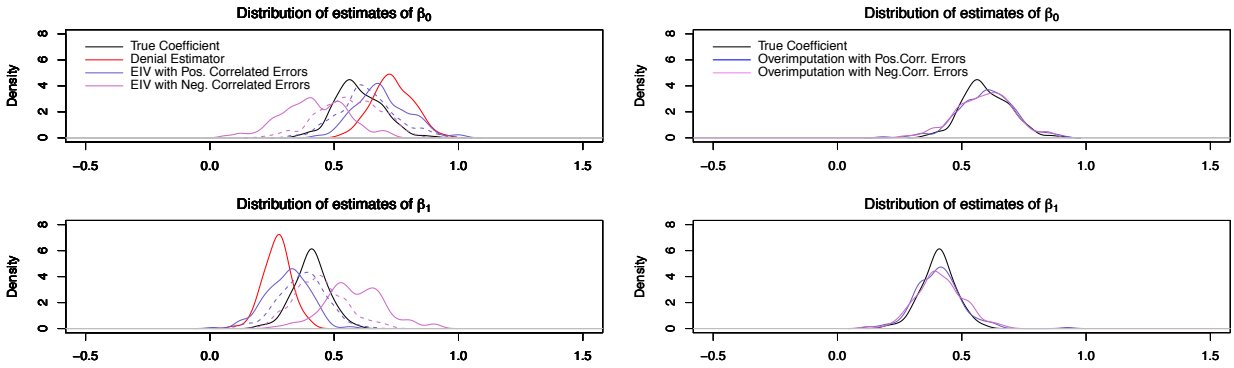
Measurement error correlated with x^*

The second requirement of the EIV model is that the measurement error is independent of the latent variable. If, for example, we believe that income is poorly measured, and wealthier respondents feel pressure to underreport their income while poorer respondents feel pressure to overreport, then the measurement error can be correlated with the latent variable.

In Figure 2.7a we demonstrate the bias this produces in EIV. Here, the error in w_2 is correlated with the latent x^*.[16] The biases are in the opposite directions as when the correlation is with y, although lesser in magnitude. Errors positively correlated with x^* lead to attenuated coefficients, and negatively correlated errors lead to overstated coefficients, as shown by the blue and violet distributions in Figure 2.7a, respectively. Dashed lines are the result of small levels of correlations (± 0.1) and the solid lines a greater degree (± 0.3).

The coefficients resulting from MO, with measurement error variance estimated from the correlation between the proxies, are contrasted in Figure 2.7b. All the distributions recover the same parameters.

16. Similar to the construction of the last simulations, we set $n = 200$, $(x^*, y, u, v) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 0, 0)$, $\Sigma = (1\ 0.4\ 0\ \rho, 0.4\ 1\ 0\ 0, 0\ 0\ \sigma_u^2\ 0, \rho\ 0\ 0\ \sigma_v^2)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$ and sequencing $\rho \in \{-0.3, -0.1, 0.1, 0.3\}$ across sets of simulations.



(**a**) Estimates from Errors-in-Variables

(**b**) Estimates from Multiple Overimputation

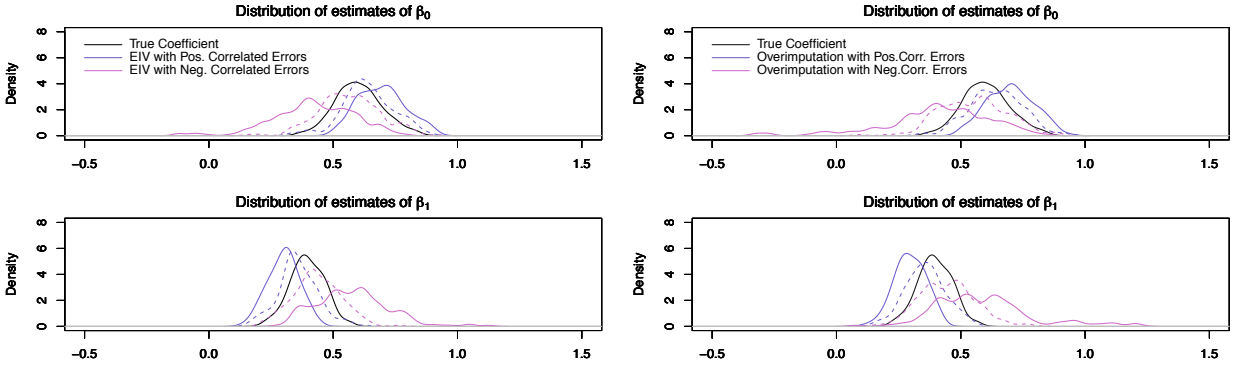
Figure 2.7: Here we show the estimates when the error in the instrument w_2 is correlated with the latent variable x^*. Positive (blue distributions) correlation leads to attentuated estimated effects in the errors-in-variables framework, and negative (violet) correlation exagerates the effect, as shown in the left. The MO estimates show no bias.

Because they sit on top of each other, only the simulations with the greatest correlation (± 0.3) are shown. For both parameters, and for both positive and negative correlation, the MO estimates reveal no bias.

Measurement errors that covary across proxies

The final condition requires the errors in the proxies be uncorrelated. If all the alternate measures of the latent variable have the same error process then the additional measures provide no additional information. For example, if we believe GDP is poorly measured, it is not enough to find two alternate measures of GDP; we also need to know that those sources are not making the same errors in their assumptions, propagating the same errors from the same raw sources, or contaminating each other's measure by each making sure their estimates are in line with other published estimates. To the extent the errors in the alternate measures are correlated, then σ_{uv} will attenuate the estimate in the same fashion as σ_u^2 did originally.

Thus, we now simulate data where the measurement errors across alternate proxies are correlated.[17] Figure 2.8a shows positively (negatively) correlated errors lead to bias in the EIV estimates that are in the same (opposite) direction as the original measurement error. Intuitively, if the errors are perfectly correlated, both the original proxy, and the alternate proxy would be the exact same variable, and thus all of the original measurement error would return. Importantly, what we see is that this is a limitation of the data that MO cannot overcome when cell level priors are directly created from the observed data. As alternate proxies contain correlated errors, identifying the amount of the variance in the proxies by the correlation of the measures is misleading. Positive or negative correlation in the measurement errors leads respectively to under or over estimation of the amount of measurement error in the data, directly biasing results as in EIV. When cell priors are set by the use of auxiliary proxies, our method continues to require the measurement errors (although not the indicates themselves of course) be uncorrelated across alternate measures, so that it is possible to consistently estimate the degree of measurement error present in the data.



(a) Estimates from Errors-in-Variables **(b)** Estimates from Multiple Overimputation

Figure 2.8: With data generated so that proxy variables have measurement error correlated with each other (so that new information is not availble with measures) both EIV (left graphs) and MO (right graphs) gives biased estimates.

17. Here we set $n = 200$, $(x^*, y, u, v) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 0, 0)$, $\Sigma = (1\ 0.4\ 0\ 0, 0.4\ 1\ 0\ 0, 0\ 0\ \sigma_u^2\ \rho, 0\ 0\ \rho\ \sigma_v^2)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$ and sequencing $\rho \in \{-0.3, -0.1, 0.1, 0.3\}$ across sets of simulations.

Even in this most difficult of settings, MO remains robust. In another set of simulations, we compare how various estimators perform when both proxies are correlated with y. Allowing these simulations to vary the amount of correlation gives an indication of how various estimators perform in this difficult situation.[18] Figure 2.9 shows that MO outperforms EIV at every level of this correlation. When the dependence between the error and y is weak, MO almost matches its zero-correlation minimum. Thus, MO appears to be robust to even moderate violations of the these assumptions, especially when compared with other measurement error approaches. Interestingly, the denial estimator can perform better than all estimators under certain conditions, yet these conditions depend heavily on the parameters of the data. If we change the effect of x^* on y from negative to positive, the performance of the denial estimator reverses itself. Since we obviously have little knowledge about all of these parameters *a priori*, the denial estimator is of little use.

Since there are gold-standard data in these simulations, we can also investigate the performance of simply discarding the mismeasured data and running MI. As expected, MI is unaffected by the degree of correlation since it disregards the correlated proxies. Yet these proxies have *some* information when the correlation is around zero and, due to this, MO outperforms MI in this region. As the correlation increases, though, it becomes clear that simply imputing the mismeasured cells has more desirable properties. Of course, with such high correlation, we might wonder if these are actually proxies in our data or simply new variables.

These simulations give key insights into how we should handle data measured with error. MO is appropriate when we have a variable that we can reasonably describe as a proxy—that is, having roughly uncorrelated, mean-zero error. Even if these assumptions fail to hold exactly, MO retains its desirable

18. These simulations follow the pattern above except they include a perfectly measured covariate, z, which determines which observations are selected for mismeasurement. Thus, we have $(x^*, y, z, u, v,) \sim N(\mu, \Sigma)$, with $\mu = (1, 1, -1, 0, 0)$ and $\Sigma = (1\ \sigma_{xy}\ -0.4\ 0\ 0, \sigma_{xy}\ 1\ -0.2\ \rho\sigma_u\ \rho\sigma_v, -0.4\ -0.2\ 1\ 0\ 0\ 0\ \rho\sigma_u\ 0\ \sigma_u^2\ 0, 0\ \rho\sigma_v\ 0\ 0\ \sigma_v^2)$ with $\sigma_u^2 = 0.5$ and $\sigma_v^2 = 0.75$. We ran simulations at both $\sigma_{xy} = 0.4$ and $\sigma_{xy} = -0.4$. Each observation had probability $\pi_i = (1 + e^{3.5+2z})^{-1}$, which has a mean of 0.25. We used the multiple proxies approach to estimating the measurement error. For EIV, we use applied the model as if the entire variable were mismeasured.

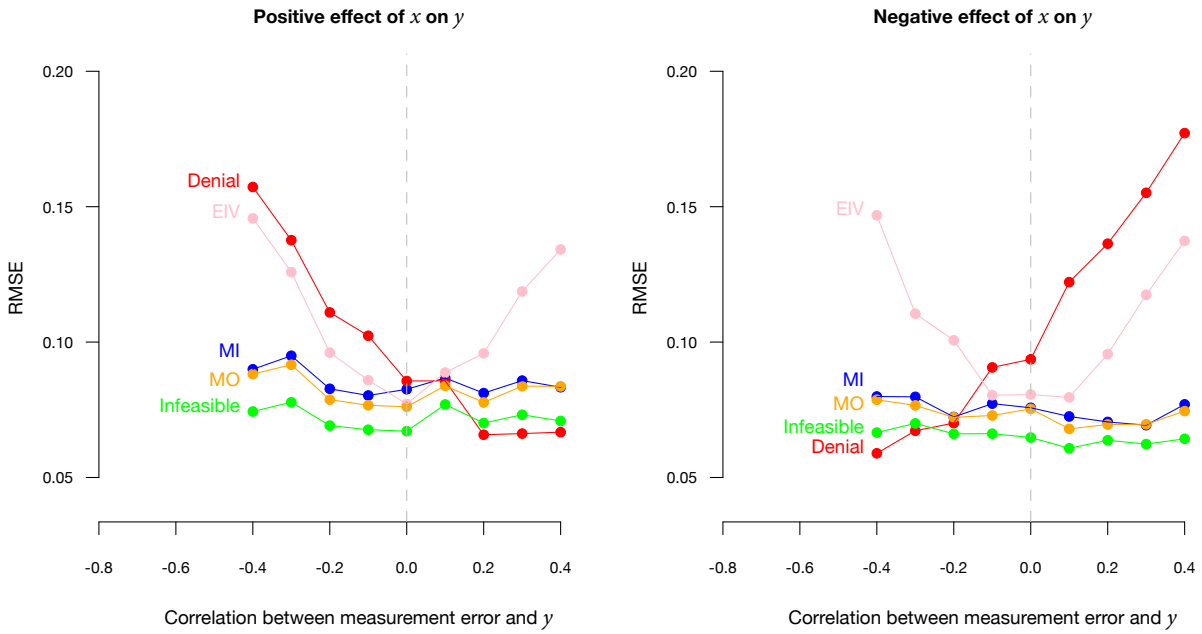


Figure 2.9: Root mean squared error for various estimators with data generated so that each proxy variable has measurement error correlated with the dependent variable. On the left, x^* has a positive relationship with y and on the right, it has a negative effect. Note that both EIV (green) and MO (orange) perform worse as the correlation moves away from zero, but MO always performs better. The denial estimator can actually perform well in certain situations, yet this depends heavily on the direction of the relationship. Both the infeasible estimator and MI are unaffected by the amount of correlation.

properties. In situations where we suspect that the measurement error on all of our proxies has moderate correlation with other variables in the data, it may be wiser to treat the mismeasurement as missingness and use multiple imputation. Of course, this approach assumes there exist gold-standard data, which may be scarce.

2.5 Empirical Applications of Overimputation

We offer three separate illustrations of the use of multiple overimputation.

2.5.1 Unemployment and Presidential Approval

To first show a practical example of the differences between our MO solution and the more common errors-in-variables (EIV) approach, we construct a measurement error process from a natural source of existing data.

It is often the case, particularly in yearly-aggregated cross-national data, that key independent variables are not measured or available at the correct point in time the model requires. Some economic and demographic statistics are only collected at intervals, sometimes as rarely as once every five or ten years. The available data, measured at the wrong period in time, is often used as a reasonable proxy for the variable's value in the desired point in time, with the understanding that there is measurement error which increases the more distant the available data is from the analyst's desired time period.

We mimic this process in actual data by intentionally selecting a covariate at increasing distance in time from the correct location, as a natural demonstration of our method in real data. In our example, we are interested in the relationship between the level of unemployment and the level of Presidential approval in the US, for which there is rich data of both series over time.[19]

We assume that the correct relationship is approximately contemporaneous. That is, the current level of unemployment is directly related to the President's approval rating. Unemployment moves over time, so the further in time our measure of unemployment is from the present moment, the weaker the proxy for the present level of unemployment, and the more the measurement error in the available data. We iteratively consider repeated models where the measurement of unemployment we use grows one additional month further from the present time.

The EIV approach relies on multiple proxies. To naturally create two proxies with increasing levels of measurement error, we use a measure of unemployment k-months before the dependent variable, and

19. Monthly national unemployment is taken from the Bureau of Labor Statistics, labor force series. Presidential approval is from the Gallup historical series, aggregated to the monthly level. We use data from 1971 to 2011. We use the last three years of each four-year Presidential term of office, to avoid approval levels within the "honeymoon" period, without adding controls into the model. We added a monthly indicator for cumulative time in office, but this only slightly strengthened these results, and so we leave the presentation as the simplest, bivariate relationship.

k-months after. That is, if we are attempting to explain current approval, we assume that the unemployment k months in the past (the k-lag) and k months in the future (the k-lead) are proxies for the current level of unemployment, which we assume is unavailable to our analyst. As k increases, the measures of unemployment may have drifted increasingly far from the present unemployment level, so both proxies employed have increased measurement error. We use these same two proxies in each of our MO models (as previously described in sections 2.2.4 and 2.3.2).

We estimate the relationship between unemployment and Presidential approval using our MO framework, and the common EIV approach, while using pairs of proxies that are from 1 to 12 months away from the present. We also estimate the relationship between approval and all individual lags and leads of unemployment; these give us all the possible denial estimators, with all the available proxies. In figure 2.10, these coefficients from the denial estimators, are shown in red, where the red bar represents the 95 percent confidence interval for the coefficient and the center point the estimated value. The x-axis measures how many months in time the covariate used in the model is from the month of the dependent variable. Positive values of x use proxies that are measured later than desired, negative values are measured too far in the past. The correct, contemporaneous relationship between unemployment and approval is in the center of this series (when x is 0) marked in black.

The EIV estimates are shown in blue. We see that with increased measurement error in the available proxies, the EIV estimates rapidly deteriorate. When the proxies for current unemployment are four months from the value of the dependent variable, the EIV estimates of the relationship are 1.40 times the true value, that is, they are biased by 40 percent. At six months the confidence interval no longer contains the true value and the bias is 98 percent. With unemployment measured at a one year gap, EIV returns an estimate 6.5 times the correct value. The MO estimates, however, are comparatively robust across these proxies. The confidence intervals expand gradually as the proxies contain less information and more measurement error. The bias is always moderate, between $+16\%$ and -12% and always clearly superior to the denial estimator, until the proxies are fully twelve months distant from the dependent

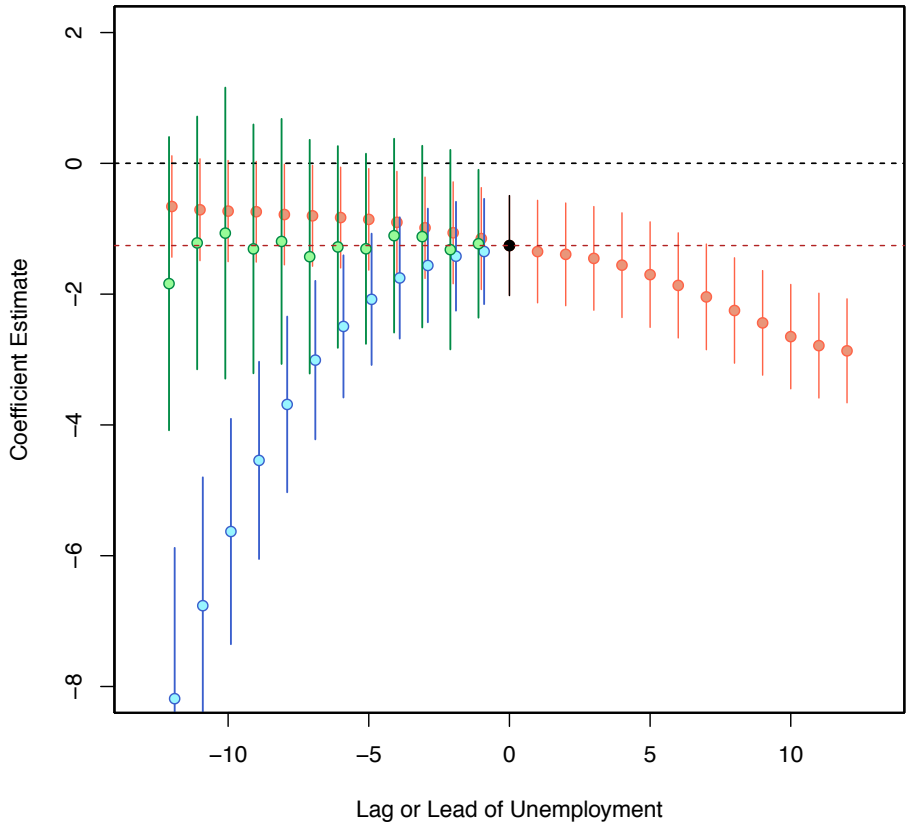


Figure 2.10: An experiment in measurement error, in the estimation of the relationship between unemployment and Presidential approval, whose true, contemporaneous value is shown in black. The blue confidence intervals represent EIV estimates of this relationship using proxies of unemployment measured increasingly distant in time. The EIV estimates fail quickly as the proxies move away from month zero. The green estimates show the robust MO estimates of the relationship. These are consistently superior to the red estimates which show the denial estimators using the unemployment rates mismeasured in time, ignoring the measurement error.

variable. Finally, at one year's distance, the MO estimates are biased by 46 percent, while the denial estimator is biased at -48 percent.

A partial explanation can be understood from our previous results. In periods where unemployment trends upwards (or downwards) the k-month lag and the k-month lead of unemployment will generally

have opposite signed measurement error. So the measurement errors in the proxies will be negatively correlated. We saw in figure 2.9 that this is a problem for both models, but that MO is much more robust to this violation than the EIV model.

We could do better than shown; we do not propose that this is the best possible model for covariates that are mismeasured in time. Adding other covariates into the imputation model could increase the efficiency of the overimputations. Averaging the two proxies would give an interpolation that might be a superior proxy to those used, and we demonstrate an application of averaging across proxies in MO in section 2.5.3. Moreover, in many applications, if there is periodic missingness over time in a variable, the best approach might be to impute all the missing values in the series with an imputation model built for time-series cross-sectional data, such as developed in Honaker and King 2010; this reinforces the main thesis of our argument, that measurement error and missing data are fundamentally the same problem. Rather, what we have shown in this example, is that in naturally occurring data, in a simple research question, where we can witness and control a measurement error process, the most commonly used model for measurement error fails catastrophically, and our framework is highly robust to even a difficult situation with proxies with negatively correlated errors.

2.5.2 Social Ties and Opinion Formation

Having looked at an example where other measurement error methodologies are available, we turn to a conceptually simple example that poses a number of difficult methodological hazards. We examine here the small area estimation challenges faced in the work of Huckfeldt, Plutzer, and Sprague (1993). The authors are interested in the social ties that shape attitudes on abortion. In particular they are interested in contrasting how differing networks and contexts, such as the neighborhood one lives in, and the church you participate in, shape political attitudes.

Seventeen neighbourhoods were chosen in South Bend, Indiana, and 1500 individuals randomly sampled across these neighborhoods. This particular analysis is restricted to the set of people who

stated they belonged to a church and could name it. The question of interest is what shapes abortion opinions, the individual level variables common in random survey designs (income, education, party identification), or the social experiences and opinions of the groups and contexts the respondent participates in. Abortion attitudes are measured by a six point scale summing how many times you respond that abortion should be legal in a set of six scenarios.

The key variable explaining abortion opinion is how liberal or conservative are the attitudes toward abortion at the church or parish to which you belong. This is measured by averaging over the abortion attitudes of *all the other people in the survey* who state they go to the same named church or parish as you mention. Obviously, in a random sample, even geographically localized, this is going to be an average over a small number of respondents. The median number is 6.[20] The number tends to be smaller among Protestants who have typically smaller congregations than Catholics who participate in generally larger parishes. In either case, the church positions are measured with a high degree of measurement error because the sample size within any church is small. This is a classic "small area estimation" problem. Here we know the sample size, mean and standard deviation of the sampled opinions from within any parish that lead to the construction of each observation of this variable.

This is an example of a variable with measurement error, where there are no other proxies available, but we can analytically calculate the observation level priors. For any individual, i, if c_i is the set of n_i respondents who belong to i's church (not including i), the priors are given by:

$$p(w_i|x_i^*) = \mathcal{N}(\bar{c}_i, sd(c_i)/\sqrt{n_i}) \tag{2.18}$$

where the $sd(c_i)$ can be calculated directly as the standard deviation within a group if n_i is generally large, or we can estimate this with the within-group variance, across all groups, as $1/n\sqrt{\sum_i (w_{ij} - \bar{w}_j)^2}$.

This is clearly a case where the measurement error is heteroskedastic; different respondents will have different numbers of fellow parishioners included in the survey. Moreover this degree of measurement

20. The mean is 10.2 with an interquartile range of 3 to 20.

error is not itself random as Catholics–who tend to have more conservative attitudes towards abortion–are from generally larger parishes, thus their church attitude will be measured with less error than Protestants who will have greater measurement error in their church attitude while being more liberal. The direction of the measurement error is still random, but the variance in the measurement error is correlated with the dependent variable. Furthermore while we have focused on the measurement error in the church attitude variable, the authors are interested in distinguishing the socializing forces of church and community, and the same small area estimation problem applies to measuring the average abortion position of the community a respondent lives in. Obviously though, the sample size within any of the 17 neighborhoods is much larger than for the parishes and thus the degree of measurement error is smaller in this variable. [21] Finally, as it is survey data, there is a variety of missing data across the variables due to nonresponse. Despite all these complicating factors this is a set up well suited to our method. The priors are analytically tractable, the heterogeneous nature of the measurement error poses no problems because we set priors individually for every cell, and measurement error across different variables poses no problems because the strength of the MI framework is handling different patterns of missingness.[22]

We replicate the final model in table 2 of Huckfeldt, Plutzer, and Sprague (1993). Our table 2.1 shows the results of the naive regression subject to measurement error in the first column. Parish attitudes have no effect on the abortion opinions of churchgoers, but individual-level variables, such as education and party identification and the frequency with which the respondent attends church predict abortion attitudes. The act of going to church seems to decrease the degree of support for legalized abortion, but

21. Within parishes, the median sample size is 6, and only 6 percent of observations have at least thirty observed responses to the abortion scale among fellow congregants in their parish. Thus we use the small sample, within-group estimate for the standard deviations, pooling variance across parishes. Within neighborhoods, however, the median sample size is 47, fully 95 percent of observations have thirty or more respondents in their neighborhood, and so we estimate the standard deviation in each neighborhood directly from only the observations in that neighborhood.

22. For additional work on small area estimation from an multiple overimputation framework, see Honaker and Plutzer 2011. In particular, there are additional possibly efficiency gains from treating the errors within individuals in the same church or community as correlated, as well as bringing in auxiliary Census data, and this work shows to approach this with two levels of imputations at both the individual and aggregated level.

	Naive Regression Model	MO Measurement Only	MO Measurement and Missingness
Constant	3.38**	-0.39	-1.68
	(1.12)	(2.09)	(1.89)
Education	0.17**	0.15**	0.14**
	(0.04)	(0.04)	(0.04)
Income	-0.05	-0.04	-0.00
	(0.05)	(0.05)	(0.05)
Party ID	-0.10*	-0.11*	-0.08*
	(0.04)	(0.04)	(0.04)
Church Attendance	-0.57**	-0.56**	-0.51**
	(0.07)	(0.07)	(0.06)
Mean Neighborhood Attitude	0.11	0.84	0.99*
	(0.21)	(0.55)	(0.48)
Mean Parish Attitude	0.13°	0.43*	0.48**
	(0.07)	(0.19)	(0.18)
Catholic	-0.48*	-0.23	-0.02
	(0.27)	(0.23)	(0.21)
n	357	521	772

** : $p < 0.01$, * : $p < 0.05$, ° : $p < 0.10$

Table 2.1: Mean Parish Attitudes are estimated by the average of across those other respondents in the survey who attend the same church. These "small area estimates" with small sample size and large standard errors have an analytically calculable measurement error. Without accounting for measurement error there is no discernable effect (column 1) but after applying MO (column 2) to correct for measurement error, we see that the average opinion in a respondent's congregation predicts their own attitude towards abortion.

the beliefs of the fellow congregants in that church have no social effect or pressure. Interestingly, Catholics appear to be different from non-Catholics, with around a half point less support for abortion on a six point scale.

The second column applies our model for measurement error, determining the observation-level priors for neighborhood and parish attitudes analytically as a function of the sample of respondents in that neighborhood and parish. Only the complete observations are used in column two, so differences with the original model are due to corrections of the measurement error in the small area estimates. We see now the effect of social ties. Respondents that go to churches where the support for legal abortion is higher, themselves have greater support for legal abortion. This may be because abortion is a moral issue that can be shaped in the church context and influenced by coreligionists, or this maybe a form of self selection of church attendance to churches that agree on the abortion issue. With either interpretation, this tie between the attitudes in the network of the respondent's church and the respondent's own personal attitude disappears due to measurement error caused by the inevitable small samples of parishioners in any individual church.

Of course our MO approach can simultaneously correct for missing data also, and multiple imputation of non-response increases by one half the number of observations available in this regression.[23] Most of the same results remain, while the standard errors shrink due to the increase in sample size. Similar to the parish variable, local neighborhood attitudes are now statistically significant at the ninety-five percent level. The one variable that changes noticeably is the dummy variable for Catholics which is halved in effect and no longer statistically significant once we correct for measurement error, and the rest of the effect disappears when we impute missing data.[24] In all, MO strengthens the author's findings, finds support for their theories in this particular model where

23. Forty-seven percent of this missingness is due to respondents who answer some, but not all, of the abortion scenarios that constitute the abortion scale. Knowing the pattern of answers to the other completed abortion questions, as well as the other control variables in the model, help predict these missing responses.

24. Catholics are still less likely to support abortion (a mean support of 3.1 compared to 3.7 for non-Catholics), but this difference is explained by variables controlled for in the model such as individual demographics and the social ties of Catholic churches which have lower mean parish attitudes than non-Catholic churches.

previously there was no result, and aligns this regression with the other models presented in their work.

2.5.3 The Effect of Political Preferences on Vote Choice

Ansolabehere, Rodden, and Snyder (2008) show that the causal effect of opinions about economic policy on vote choice is much stronger than previously estimated (but consistent with what one would expect) via a simple alternative method of removing measurement error: averaging many multiple measures of the same concept. Although the data requirements make approach only occasionally applicable, it is powerful, when possible, and instructive. They consider $K = 34$ survey items $\{w_1, w_2, \ldots, w_K\}$, all taken to be imperfect indicators of an unobserved variable, x, and assume common measurement error variance σ_x^2. That is, $w_{ik} = x_i + u_{ik}$ for each i, where $E[u_{ik}] = 0$ and $E[u_{ik}^2] = \sigma_k^2$. While any individual measure has variance $\sigma_x^2 + \sigma_k^2$, the average of the measures, $\bar{w}_i = \frac{1}{K}\sum_{k=1}^{K} w_{ik}$ has variance $\sigma_x^2 + \bar{\sigma}^2/K$, where $\bar{\sigma}^2 = \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2$ is the average measurement error variance among the items. If all of the measures have similar amounts of measurement error, then the average of the items will have far lower levels of measurement error than any single item. Furthermore, the effect of measurement error will decrease as K increases.

We now show that in the more usual situation where researchers have access to one or only a few measures of their key concepts, MO can still recover reliable estimates because it makes more efficient use of the data and available assumptions. It also provides more information by enabling one to avoid the assumption that all available measures are indicators of exactly the same underlying concept.

To illustrate these features, we reanalyze Ansolabehere, Rodden, and Snyder (2008) with their data from the American National Election Study. Using their general approach, we find that a one standard deviation increase in economic conservatism leads to an 0.24 increase in the probability of voting for Bob Dole.

We then perform MO using only two of the thirty-four variables. To avoid cherry picking results, we reran the analysis using all possible subsets of two variables chosen from the available 34. For each of

these pairs, we overimputed the first variable, using the second as a proxy (see Section 2.3.2). We then estimate the effect of that overimputed variable on voting for Bob Dole using a probit model. We compare this method with simply taking the pairwise averages and using them as the measure of economic policy preferences. These approaches mimic a common situaton in political science when researchers have access to relatively few variables.

Figure 2.11 shows the relationship between the two estimates. Each column represents the average of the estimated effects for one measure, averaged across all its pairs. Note that for every variable, MO estimates a larger effect than does averaging, as can be seen by the positive slope of every line. The "gold-standard" estimate suggested by Ansolabehere, Rodden, and Snyder (2008) is well above the any of the pairwise averaging estimates, but it lies firmly in the middle of the pairwise MO estimates. This striking result shows that MO makes more efficient use of the available data to correct for measurement error.

While the average results of the pairwise MO align with the thirty-four measure gold-standard, there is considerable variance among the individual measures. This is in part due to a fundamental difference between MO and averaging (or more general scale construction techniques like factor analysis). MO corrects measurement error on a given variable instead of constructing an new measure of an underlying concept. This often valuable result allows us to investigate how the estimated effect of economic preferences varies across the choice of measure. With pairwise MO, we find that classic economic ideology items regarding the size of government and its role in the economy have a much larger estimated effect on vote choice than questions on welfare policy, equal opportunity, and poor people — all of which were treated the same under averaging. It is interesting to note that correcting the classic questions on economic policy lead to even higher estimates than implied by the gold-standard. Furthermore, the lowest estimated effects come from variables that relate to views of the poor and their benefits from the government, which in part may proxy for other issues such as racial politics.

As Ansolabehere, Rodden, and Snyder (2008) point out, averaging is a "tried and true" method for

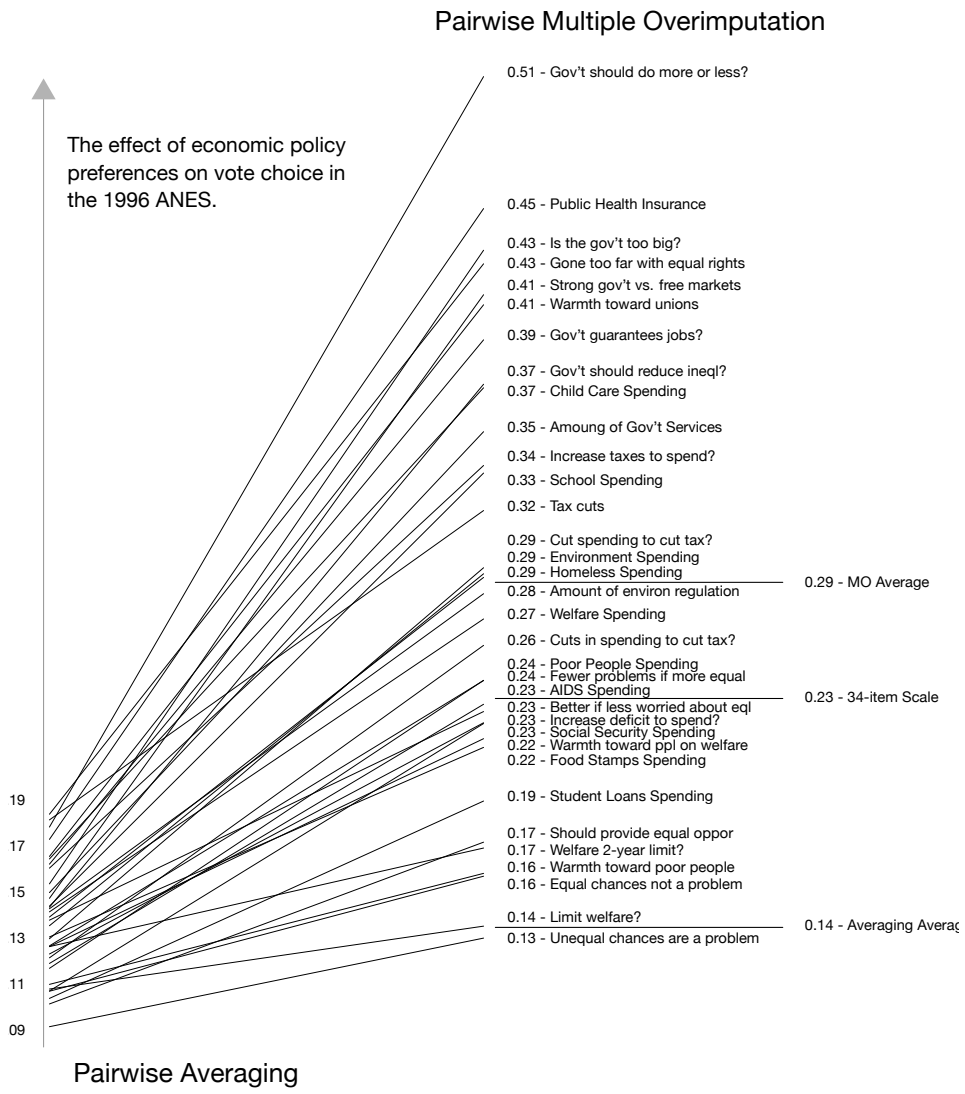


Figure 2.11: The lines connect estimates from averaging across all pairwise estimates containing the specified variable (left) and estimates from multiple overimputation (right). MO estimates a higher average effect, and one that is closer to the "gold-standard" 34-item scale in each case. Furthermore, MO finds higher estimated effects for classic economic ideology questions and lower effects for questions on welfare and economic opportunity.

alleviating measurement error and it works very well when a battery of questions exists for a given concept. When, as usual, less information is available, MO may be able to extract more information from the available data.

2.6 What Can Go Wrong?

MO's 2-step estimation procedure makes it, like MI, highly robust to misspecification, especially compared to structural equation-like approaches. However, like any statistical procedure, using it inappropriately can lead to incorrect inferences. Inappropriate uses include the following. First, using MO, or any measurement error correction procedure, to deal with very small degrees of measurement error may reduce bias at the expense of a larger increase in variability. Given the likely high levels of measurement error in political science variables, this is a concern, but will not normally be much of an issue. Second, overestimating the amount of measurement error in an MO application can lead to incorrect inferences, but these inferences will typically remain within the minimal-assumption bounds, and so users should be sure to consult the bounds as a check. Further, MO handles these situations better than, say, method-of-moments estimators (Fuller, 1987) even in simple cases.[25] Third, violations of the key assumptions about measurement error, especially MAR assumptions, can lead to bias and, like assumptions about omitted variable bias or ignorability, are not normally testable without additional data. Sensitivity tests could be conducted however. Finally with these qualifications, there are conditions under which simple techniques like listwise deletion or ignoring the problem altogether will be preferred over MO, but these conditions normally make it highly unlikely that one would continue to trust the data for subsequent analyses (King et al. 2001).

25. In the simulations of Section 2.3.1, we find that a method-of-moments estimator can have up to 188 times higher squared bias without any offsetting increase in efficiency.

2.7 Conclusion

Measurement error is a common, and commonly ignored, problem in the social sciences. Few of the methods proposed for it have been widely used, largely because of implausible assumptions, high levels of model dependence, difficult computation, and inapplicability with multiple mismeasured variables.

Here, we generalize the multiple imputation framework to handle observed data measured with error. Our multiple overimputation (MO) generalization overwrites observed but mismeasured observations with a distribution of values reflecting the best guess and uncertainty in the latent variable. Our conceptualization of the problem is that missing values are merely an extreme form of measurement error, and in fact an easy case to address with standard imputation methods because there is so little to condition on in the model. However, correctly implementing the multiple imputation framework to also handle "partially missing" data, via informative observation-level priors derived from the mismeasured data, allows us to unify the treatment of all levels of measurement error including the case of completely missing values.

This approach makes feasible rigorous treatment of measurement error across multiple covariates, with heteroskedastic errors, and in the presence of violations of assumptions necessary for common measurement treatments, such as the errors-in-variables model. The model works in survey data and time series, cross-sectional data, and with priors on individual missing cell values or those measured with error. With MO, Scholars can preprocess their data to account for measurement error and missing data, and then use the overimputed data sets our software produces with whatever model they would have used without it, ignoring the measurement issues. These advances, along with the more application-specific techniques of Imai and Yamamoto (2010) and Katz and Katz (2010), represent important steps for the correction of measurement error in the social sciences.

The advances described here can be implemented when the degree of measurement error can be analytically determined from known sample properties, estimated with additional proxies, or even when it can only be bounded by the analyst. However, looking forward, often the original creators of

new measures are in the best position to know the degree of measurement error present (for example, through measures of intercoder reliability, comparison to gold-standard validation checks, or other internal knowledge) and we would encourage those who create data to include their estimates of variable or cell-level measurement error as important auxiliary information, much as sampling frame weights are considered essential in the survey literatures. Now that easy-to-use procedures exist for analyzing these data, we hope this information will be made more widely available and used.

A change is gonna come.

Sam Cooke

3

Game-changers: Detecting Shifts in the Flow of Campaign Contributions

3.1 Introduction

Electoral campaigns are the central events in the political life of democracies. And, increasingly, campaigns are as much about garnering money as they are about garnering votes. Indeed, candidates view fundraising as a vital and time-consuming part of what they do. For citizens, campaign contributions represent a costly form of political participation. This participation certainly depends on

features of the individual (Verba, Schlozman, and Brady 1995), but it also ebbs and flows throughout the election season in response to news coverage, campaign events, and changes in candidate strategy (Mutz 1995). While there is some empirical evidence that momentum matters (Bartels 1985), there have been few studies that attempt to pinpoint statistically *when* campaigns take off or fall flat. This essay seeks to do just that: find points in time when contributions to a candidate change dramatically.

To estimate these shifts, I propose a novel Bayesian changepoint model, called the *game-changers* model, tailored to handle campaign contribution data. The number of contributors to a campaign on a given day is highly overdispersed due to the clustered processing of contributions and the intermittent nature of political attention. Extant changepoint models for count data, such as those used in political science (Park 2010; Spirling 2007), use the Poisson distribution, which is problematic here because it places inappropriate restrictions on the variance of the data. As shown below, this can lead to incorrect inferences on the location of changepoints. The game-changers model uses random effects to handle overdispersion, an approach that is equivalent to assuming a negative binomial likelihood, which is common in political science (King 1989).

Most changepoint models require researchers to specify the number of changepoints in advance, but it is hardly clear what the "correct" number of game-changers is for any given campaign, let alone a series of campaigns. To alleviate this problem, the game-changers model takes a Bayesian nonparametric approach and estimates the number of changepoints along with their location. This approach is an extension of the Chib (1998) method for estimating changepoints and incorporates a Dirichlet process prior for the clustering of the contributions into regimes. This provides a computationally efficient and conceptually straightforward method for allowing the model to include the number of regimes and, thus, the number of changepoints. While this model is tailored to estimating changepoints for campaign contributions, it can be applied to any time series of overdispersed counts. More generally, the Dirichlet process prior approach to estimating the number of changepoints generalizes beyond this model or even count data.

The essay proceeds as follows. Section 3.2 describes the campaign contributions data and the various factors that lead to overdispersion. Section 3.3 describes the game-changers model and the computational approach to fitting the model. Section 3.4 describes four vignettes that show how the model works in simulated and real data. Section 3.5 concludes.

3.2 The dynamics of campaign contributions

The Federal Election Commission (FEC) collects data on contributions of $200 or more to campaigns for federal office made by individuals and groups. The FEC requires campaigns to report a fair amount of information, including the date that the campaign received the contribution (Federal Election Commission 2011). These reports allow us to track both the daily number of contributions made to a campaign along with the amount contributed.

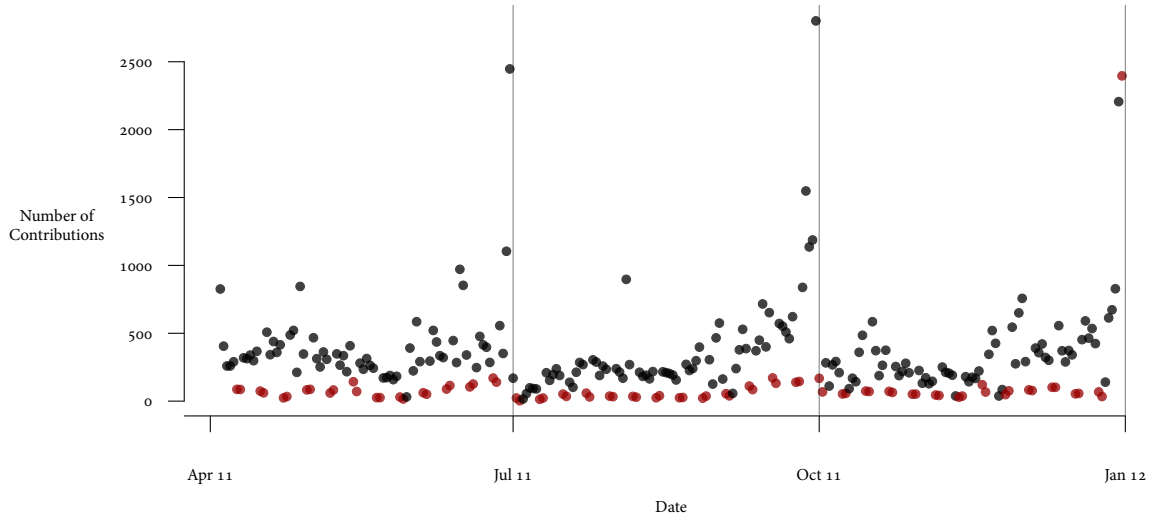


Figure 3.1: Daily number of individual contributions to Barack Obama in 2011. Black dots are weekdays and red dots are weekends. Vertical lines are the filing deadlines.

Campaign contributions have a few unique features that make it difficult to apply commonly-used

changepoint models. For non-electronic contributions, the date the campaign reports "receiving" a contribution might differ from the date that the date that the donor made the contribution. There are many reasons for this, but two stand out. First, if contributions travel by mail, they might take some time to reach the campaign. Second, and more important for this study, is the potential for delays in campaign processing of contributions. Campaign staff are limited in the amount of time they can process incoming contributions–even if a contribution arrives on a given day, it might not be processed until later. One indication of this is given in Figure 3.1, which shows the number of contributions received by Barack Obama in 2011, with weekends plotted in red. Campaigns are much more likely to receive contributions on a weekday compared to the weekend. This pattern results from the fact that campaign staff largely work a traditional work week and so contributions that arrive during the weekend are processed after staffers return to work on Monday.

Candidates also have strategic reasons for processing contributions at different rates, one due to signaling and one due to contribution limits (Christenson and Smidt 2011). First, the FEC requires that candidates report their contributions to the FEC at various points throughout the campaign. These reports are important as they publicly disclose the candidate's ability to raise funds. Candidates want to signal that they are a high-quality candidate and one way to do this is to have a large number of contributors. Thus, campaign staff work to process any incoming contributions before these filing deadlines so as to maximize the reported contributions. Figure 3.1 shows the filing deadlines for 2011 as vertical lines. Clearly, there is a marked increase in the number of contributions received around the filing deadlines. A second reason for pre-deadline increases is that there are different contribution limits for before and after the primary election. A candidate would want to process any pre-primary contributions before the relevant filing deadline so that those pre-primary donors can legally contribute to the campaign again during the run up to the general election. Filing deadlines and weekends are two features campaign contributions data that contribute to the overdispersion of their distribution.

The above reasons for overdispersion could be measured and accounted for in a Poisson regression

model, which would alleviate some of the problem. There are other features of campaign contributions that can lead to overdispersion as well, some of which are hard to measure. For instance, campaigns receive many contributions as part of campaign fundraising events—dinners, speaking engagements, and so on. These events add to the clustering of the contributions because they group contributors together in time. These events are more problematic than weekends and filing deadlines because it is very difficult to collect data on the timing of campaign fundraisers. Thus, it is important that we build a model that can handle these unmeasured forms of overdispersion inherent in contributions data.

3.3 A model for changepoints in campaign contributions

3.3.1 Changepoint models

Changepoint models estimate discrete changes in the distribution of time-series data. Given a time-series of observed contribution counts, $Y = (y_1, \ldots, y_T)$, a changepoint model assumes that the distribution of y_t is distributed according to a parameter γ_t, which takes on $M + 1$ distinct values, depending on t, where M is the number of changepoints. Thus, $M + 1$ is the number of distinct parameter regimes in the data. Let $\mathbf{c} = (c_1, \ldots, c_M)$ be the vector of changepoints and $\theta = (\theta_1, \ldots, \theta_{M+1})$ be the vector of parameters associated with each regime. With these, we can define the parameters at each point in time as

$$\gamma_t = \theta_m \qquad \text{i.f.f.} \qquad c_{m-1} < t \leq c_m, \tag{3.1}$$

where we define $c_0 = 0$ and $c_{M+1} = T$. Thus, each observation takes on the parameters of its regime.

One way to conceptualize this model is to imagine the time-series as residing in one regime for a given amount of time before jumping to another regime at a changepoint. Chib (1998) shows that we can think of this regime-switching structure as a discrete-time, discrete-state Markov process with a constrained transition matrix. Let $S = (s_1, \ldots, s_T)$ be a vector of regime indicator, so that if $s_t = m$,

then at time t the time-series is in regime m and that $c_{m-1} < t \leq c_m$. Given the nature of the model, we only have to specify the probability of transitioning to the next regime: $\Pr(s_{t+1} = j+1 | s_t = j) = p_{j,j+1}$. We can model S in place of the changepoints since the kth changepoint happens at c_k if and only if $s_{c_k} = k$ and $s_{c_k+1} = k+1$. The regime indicators are useful in Bayesian changepoint models, where we can augment a model with these latent variables to ease computation (Chib and Greenberg 1996). This model of Chib (1998) forces the time-series to reside in each of the $M+1$ regimes without skipping a regime or returning to a regime. Note though, that if we are interested in estimating the changepoints, $\mathbf{c}$, then recurrent regimes are straightforward since the model will recover the relevant changepoints and treat these recurrence as distinct regimes. More troubling is the lack of regime skipping, which means that each of the $M+1$ regimes is visited. This can be problematic if the true number of structural breaks is less than the number of changepoints in the model. I address this issue below.

3.3.2 Tailoring changepoints for campaign contributions data

Up to this point, I have left the distribution of y_t unspecified since changepoint models can accommodate many different data-generating processes, including continuous, binary, and count outcomes. See Park (2010, 2011) and Spirling (2007) for different applications of changepoint models in political science. Unfortunately, the extant changepoint models are poorly suited to handle campaign contributions data due to the features discussed above.

The overdispersion inherent in campaign contributions data requires a deviation from the Poisson changepoint models of Chib (1998), Park (2010), and Spirling (2007). These models assume that

$$y_t | \lambda_t, s_t = k \sim \text{Po}(\lambda_t), \qquad \lambda_t = \exp(X_t \beta_k), \tag{3.2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{M+1})$ are the Poisson regression coefficients from each regime. Given the nature of the Poisson distribution, these models implicitly assume that the mean in any specific regime is equal to the variance. This assumption is unlikely to hold in general and fails miserably in campaign

contributions data (see Section 3.4.1 for a demonstration of this).

As shown by Frühwirth-Schnatter et al. (2009) in the context of mixture modeling, we can handle overdispersion in a count model by augmenting model 3.3 with a random intercept:

$$y_t|\lambda_t, \eta_t, s_t = k \sim \text{Po}(\eta_t\lambda_t), \qquad \lambda_t = \exp(X_t\beta_k). \tag{3.3}$$

The random effects, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_T)$, allow for the marginal distribution of the data (that is, $p(y_t|\lambda_t)$) to have a separate mean and variance. In fact, if we place a Gamma prior on the random intercept,

$$\eta_t|s_t = k, \rho_k \sim \text{Ga}(\rho_k, \rho_k), \tag{3.4}$$

then the marginal distribution of the data is negative binomial. Note that the prior in (3.4) allows for different amounts of overdispersion in different regimes. As ρ_k tends toward infinity, the model converges to a Poisson model. For a given finite value of ρ_k, the marginal distribution of the data has the following form:

$$p(y_t|\lambda_t, \rho_k, s_t = k) = \binom{\rho_k + y_t - 1}{\rho_k - 1} \left(\frac{\rho_k}{\rho_k + \lambda_t}\right)^{\rho_k} \left(\frac{\lambda_t}{\rho_k + \lambda_t}\right)^{y_t}, \tag{3.5}$$

which is a negative binomial with trial size ρ_k and probability of success $\rho_k/(\rho_k + \lambda_t)$. Negative binomial models are common in political science for handling count data with overdispersion (King 1989).

3.3.3 Estimating the number of changepoints

In order to estimate the location of the changepoints, most existing changepoint models require we know the *number* of changepoints that exist in the data. Obviously, for almost any campaign, it would extraordinarily difficult for researchers to know, with certainty, the number of changepoints in the data. For most researchers, in fact, estimating the number of changepoints might be as interesting as

estimating their location. A common approach in changepoint models is to estimate many models, each conditional on a number of changepoints, then use a model selection tool to choose the "best" model (Park 2010; Chib 1998).

Changepoint models, though, are a special type of finite-mixture model and these types of models fail to meet the regularity conditions of the traditional, likelihood-based non-nested model comparison tests. Therefore, a common way to compare models is to use Bayesian model selection via the calculation of the marginal likelihood of the model. Park (2011) provides an example of how this approach works for binary and ordinal-probit changepoint models. Chib (1995) provides a straightforward approach to calculating marginal likelihoods when using MCMC based on the Gibbs sampler. This approach is not applicable with the above negative binomial model, however, because it requires a Metropolis-Hastings step to draw the ρ_k. Alternative approaches to Bayesian model comparison are computationally difficult and pose problems with highly unlikely models (Park 2011, p. 192). Koop and Potter (2009) identify another major problem with fixed in-sample changepoint approaches: common Bayesian priors, such as those used in Chib (1998), lead to undesirable behavior at the end of the sample.

An alternative to model selection is to estimate the number of changepoints as part of the model itself. A number of methods have been proposed to leave the number of changepoints unrestricted, but many of these approaches are based on a conditionally linear model and not appropriate for the above non-linear model.[1] Instead, this essay preserves the simplicity and computational efficiency of the method proposed by Chib (1998) but allows it to choose the number of changepoints as part of the model.

The approach of Chib (1998) assumes that there are $M + 1$ regimes and that each of these regimes is visited by the time-series. The model imposes this restriction by assuming that $s_1 = 1$ and that

1. Giordani and Kohn (2008) provide a method of estimating the number of changepoints that work for conditionally linear, Gaussian models. Geweke and Jiang (2011) and Chong and Ko (2011) provide alternative MCMC implementations of process priors in changepoint models. Koop and Potter (2007) amends Chib's method to allow for the estimation of the number of regimes, but this approach requires many more calculations than the present approach.

$s_T = M + 1$. This essay instead places no restriction on the value s_T, so that the model can estimate fewer than $M + 1$ regimes in the observed sample. This shifts M from being the assumed number of changepoints to the maximum number of changepoints allowed by the model. This approach will recover the posterior distribution on the number of changepoints, as long as we set M high enough not to truncate the posterior. Note that we can only observe T possible regimes in the data—one for each observation.

We can represent this approach as using a specific version of the Dirichlet process prior, a popular tool in Bayesian nonparametrics (Neal 2000). The Dirichlet process prior creates an *infinite* mixture model as opposed to the *finite* mixture models that are typically used by changepoint models.[2] In general, models with Dirichlet process priors group observations together into a countably infinite set of groups (Ferguson 1973; Escobar and West 1995). We can show the central intuition of the Dirichlet process prior as by taking the limit of finite mixture models. Suppose we have a mixture model with the same models as above and K components:

$$y_t | s_t, \boldsymbol{\beta}, \boldsymbol{\rho}, \eta_t \sim \text{Po}(\eta_t \exp(X_t \beta_{s_t})) \tag{3.6}$$

$$s_t | \mathbf{p} \sim \text{Discrete}(p_1, \ldots, p_K) \tag{3.7}$$

$$(\beta_k, \rho_k) \sim G_0 \tag{3.8}$$

$$\mathbf{p} \sim \text{Dirichlet}(b/K, \ldots, b/K). \tag{3.9}$$

Here, G_0 is the "base" distribution of the regime parameters. Neal (2000) shows that we can marginalize over the distribution of $\mathbf{p}$ and, as $K \to \infty$, we find that:

$$p(s_t = k | s_1, \ldots, s_{t-1}) \to \frac{n_{t,k}}{t - 1 + b} \tag{3.10}$$

$$p(s_t \neq s_j \text{ for all } j < t | s_1, \ldots, s_{t-1}) \to \frac{b}{t - 1 + b} \tag{3.11}$$

2. For other uses of Dirichlet process priors in political science, see Grimmer (2011) and Spirling and Quinn (2010).

Here $n_{t,k}$ is the number of observations up to time t are in component k. Thus, each observation is allocated to a component with a probability that is proportional to the number of previous units already allocated to that component. This property of the Dirichlet process prior is called the "rich get richer" property and is a fundamental assumption of the prior. Different Bayesian nonparametric priors have different assumptions embedded into their design and these different assumptions can lead to different clusterings. With this prior in hand, Neal (2000) provides a host of MCMC algorithms to estimate the posterior distribution of both the clusters and the cluster parameters.

A changepoint model is a special case of a clustering, where we refer to the clusters as regimes and restrict how the observations move from regime to regime. Namely, we stipulate that an observation at time t must either be in the same regime as observation $t-1$ or it can form a new regime. Observations cannot "return" to a previous regime. Thus, the mixing probabilities $\mathbf{p}$ do not follow the symmetric Dirichlet distribution of (3.9). For s_{t+1}, all p_k are 0 with the exception of p_{s_t} and $p_{s_t+1} = 1 - p_{s_t}$. These are the probability of remaining in the same regime as t and the probability of moving to a new regime. Since there are only two possibilities, our prior over these values becomes a Beta distribution with parameters a and b. This setup implies a Dirichlet process prior with the following transition probabilities as $K \to \infty$:

$$p(s_t = k | s_{t-1} = k, s_1, \ldots, s_{t-2}) \to \frac{n_{t,k} + a}{n_{t,k} + a + b} \tag{3.12}$$

$$p(s_t = k+1 | s_{t-1} = k, s_1, \ldots, s_{t-2}) \to \frac{b}{n_{t,k} + a + b}. \tag{3.13}$$

Note that these transition probabilities are no longer Markovian, as they are in the original Chib (1998). This only requires a modest adjustment to the algorithm to draw the s_t.

In practice, there is no need to draw parameters for an infinite number of regimes. Instead of sampling from the infinite mixture model, I take an alternative approach that uses a truncated approximating distribution with a finite, but large, number of regimes (Ishwaran and James 2001). This

will not limit the number of regimes estimated by the model, so long as the upper bound on the number of regimes is large enough to never truncate the distribution in practice. In the empirical examples below, I use an upper bound of 20 changepoints and there is never more than 11 changepoints estimated in any iteration of the MCMC algorithm.

3.3.4 Priors and hyperparameters

The complete model requires proper priors on all parameters and I use the following:

$$\rho_k \propto \rho_k^{e-1}(\rho_k + d)^{e+f}; \tag{3.14}$$

$$\beta_k \sim \mathcal{N}(0, B_0); \tag{3.15}$$

$$p_{k,k+1} \sim \text{Beta}(a, b). \tag{3.16}$$

The prior for each regime parameters are *a priori* independent. In order for the posterior to exist, the priors must be proper, which means that $e > 1$ for the prior on ρ_k. For all of the models below, I use $e = f = 2$ and $d = 10$, which follows Frühwirth-Schnatter et al. (2009), and $B_0 = 100$.

The priors on $p_{k,k+1}$ imply a prior on the length of each regime and, therefore, a prior on the number of regimes that are visited in the sample. Namely, $p_{k,k+1}$ is the probability of a one-period regime, which we can build up to infer an expected *a priori* regime length. In the applications below, I use $a = 20$ and $b = 0.1$, which implies an expected regime length of 200 days and around 1.5 regimes observed in a typical election season. These priors are intentionally designed to allow for long regimes and potentially no changepoints at the expense of finding shorter regimes. When we assume shorter regimes *a priori*, we end up identifying clusters of one- or two-day outliers in addition to the more clearly "game-changing" changepoints. In any case, the estimated changepoints do not vary too much as we change the value of the hyperparameters a and b.

3.3.5 A Markov chain Monte Carlo estimation strategy

Given the above model, we can write the posterior as follows:

$$
\begin{aligned}
p(\mathbf{s}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\eta} | \mathbf{y}, \mathbf{X}) \propto\ & p(y_1|\beta_1, \eta_1) p(\eta_1|\rho_1) \times \\
& \prod_{t=2}^{T} \left[\sum_{m=1}^{M+1} p(y_t|\beta_m, \eta_t) p(\eta_t|\rho_m) p(s_t = m|\beta_m, \rho_m) \right] \times \\
& \prod_{i=1}^{M+1} p(\beta_m|B_0) p(\rho_m|d, e, f) p(p_{i,i+1}|a, b)
\end{aligned} \tag{3.17}
$$

To sample from this, I take a Markov chain Monte Carlo approach using Gibbs sampler which samples from the full conditional posterior of each parameter. Below, I discuss the non-standard steps in detail.

Drawing the latent regimes

To draw the latent states, I use a modified version of the Chib (1998) algorithm. Chib points out that we can write the full conditional posterior of $\mathbf{s}$ as

$$
p(s_T|\mathbf{y}, \Theta, P) \times p(s_{T-1}|\mathbf{y}, s_T, \Theta, P) \times \cdots \times p(s_t|\mathbf{y}, \mathbf{s}_{t+1:T}, \Theta, P) \times \cdots \times p(s_1|\mathbf{y}, \mathbf{s}, \Theta, P), \tag{3.18}
$$

where $\Theta = (\boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\eta})$ is the collection of the model parameters, $P = (p_{1,2}, \ldots, p_{M,M+1})$ is the collection of transition probabilities, and $\mathbf{s}_{t+1:T} = (s_{t+1}, \ldots, s_T)$. Crucially, note that Chib (1998) drops the term for s_T because Chib assumes the last observation is in the last regime, $M + 1$, with probability one. In this specification, we allow s_T to take any value between 1 and $M + 1$, with a probability determined by the data. With this in hand, we can derive each of these distribution and then sample from each, in turn:

- s_T from $p(s_T|\mathbf{y}, \Theta, P)$,

- s_{T-1} from $p(s_{T-1}|\mathbf{y}, s_T, \Theta, P)$,
- $\vdots$
- s_2 from $p(s_2|\mathbf{y}, \mathbf{s}_{3:T}, \Theta, P)$.

The regime of the first period is always $s_1 = 1$. Thus, to sample from this, it is sufficient to sample from $p(s_t|\mathbf{y}, \mathbf{s}_{t+1:T}, \Theta, P)$, which is given by Chib (1998).

Drawing the model parameters

Now that we have draws of the latent states, we need to take draws of the model parameters in each regime (β_k, ρ_k). The non-linear nature of the distributions involved eliminate the possibility of closed-form posterior distributions. This makes the straightforward application of Gibbs sampling impossible. To avoid the inefficiencies of other MCMC approaches, I draw on the auxiliary mixture sampling approach of Frühwirth-Schnatter et al. (2009). This approach augments the data with a set of latent variables τ_{t1} and τ_{t2} which contain all the distributional information about the outcome y and whose distribution can be approximated by a mixture of Normals. With draws of $\tau_t = (\tau_{t1}, \tau_{t2})$ and mixture component indicators $r_t = (r_{t1}, r_{t2})$, we can turn this non-linear problem into a linear Gaussian regression problem. That is, conditional on τ_t, r_t, and η_t, posterior inference on the β_k is simply a Bayesian linear regression. Frühwirth-Schnatter et al. (2009) also shows how to include draws for the negative binomial parameters ρ_k and η_t in a Gibbs sampler.

MCMC algorithm

Thus, I proceed to draw from the posterior using the following Gibbs sampling approach:

1. Draw $\mathbf{s}|\mathbf{y}, \Theta, P$ as above.
2. Draw $(\boldsymbol{\rho}, \boldsymbol{\eta})|\mathbf{y}, \boldsymbol{\beta}, \mathbf{s}$:

(a) Draw $\rho_k|\mathbf{y}, \boldsymbol{\beta}$ unconditional on $\boldsymbol{\eta}$ using a Metropolis-Hastings step.

(b) Draw $\eta_t|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\rho}, \mathbf{s} \sim \text{Gamma}\left(\rho_{s_t} + y_t, \rho_{s_t} + \exp(X_t\beta_{s_t})\right)$, for $t = 1, \ldots, T$.

3. Sample $\boldsymbol{\tau}, \mathbf{r}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\rho}$ using the auxiliary mixture approach Frühwirth-Schnatter et al. (2009).

4. Draw from $\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{r}, \boldsymbol{\eta}$ using the auxiliary mixture approach of Frühwirth-Schnatter et al. (2009).

5. Draw $p_{ii}|\mathbf{s}, a, b$ from $\text{Beta}(a + n_{ii}, b + 1)$, for $i = 1, \ldots, M + 1$.

3.4 Vignettes

3.4.1 A simulation study

To demonstrate the effectiveness of the gamechangers model, I apply it to a simulated dataset, seen in the bottom panel of Figure 3.2. This dataset has $T = 200$ observations with four regimes with 50 observations each. I simulated the data in each regime with a simple intercept, so that $\boldsymbol{\beta} = (6, 3, 6, 3)$ and with overdispersion parameters $\boldsymbol{\rho} = (1.5, 0.5, 3, 1.5)$. I ran the above MCMC sampler with an upper bound of 20 changepoints for 10,000 iterations, thinned by 10, with a burin period of 5,000 iterations.

The nonparametric nature of the sampler makes visualizing the posterior more complicated than in more traditional approaches to changepoint problems. Namely, since the number of regimes can change from iteration to iteration, it makes little sense to look at the probability of a given observation residing in a specific regime—the nature of the regimes themselves are changing. An alternative approach is to simply calculate the *posterior changepoint probability*, which is simply

$$\hat{c}_t = \frac{1}{G}\sum_{g=1}^{G} \mathbb{I}(\hat{s}_t^{(g)} = j + 1, \hat{s}_{t-1}^{(g)} = j), \tag{3.19}$$

where $\mathbb{I}()$ is an indicator function and $s_t^{(g)}$ is the gth draw of the regime for observation t. We can calculate this straightforwardly from the MCMC output by finding the proportion of draw where a

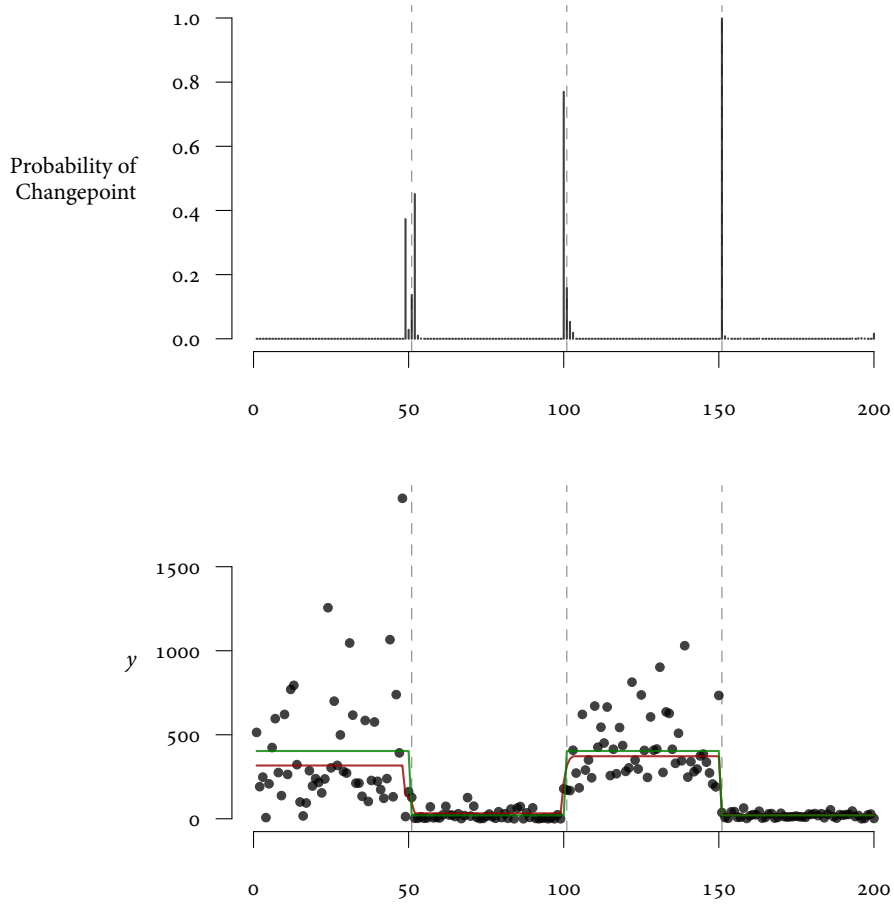


Figure 3.2: Changepoints in a simulated example. On top, the posterior probability of a changepoint in a given period. On bottom, the simulated data with the true daily mean (green), posterior mean (blue), and true changepoints (vertical and dashed).

change occurs at t. The top panel of Figure 3.2 shows these values for the simulated data. It is clear that there is a high posterior probability of the changepoints occurring around their true values of $t = 51$, $t = 101$, and $t = 151$.

Making inferences about the regime parameters is also difficult due to the changing number of regimes. Regime 2 in one draw could be very different from regime 2 in another draw. Instead of investigating the posterior mean of the regime parameters, we can estimate the posterior mean of the

observation. That is, we can estimate

$$\hat{\lambda}_t = \frac{1}{G} \sum_{g=1}^{G} \exp(X_t \hat{\beta}_{s_t^{(g)}}^{(g)}), \tag{3.20}$$

where G is the number of MCMC draws, $\hat{\beta}_k^{(g)}$ is the draw of β_k in iteration g of the sampler. The bottom panel of Firgure 3.2 overlays the true values of λ_t in green along with its posterior mean, $\hat{\lambda}_t$ in red. In this case, the posterior values largely matched up the truth, with some (small) shrinkage toward the prior.

Extant changepoint models in political science also rely on the Poisson distribution, but do not take into account overdispersion. To demonstrate this, I applied the Poisson changepoint model of Park (2010) to the same set of simulated data. For this model, we must specify the number of changepoints, so to give an advantage, I correctly specify the number of changepoints. Even with this advantage, the Poisson model is unable to recover the true locations of the changepoints. The top panel of Figure 3.3 shows that none of the estimated changepoints come close to the true changepoints. The bottom panel of the same figure shows why the Poisson model fails to find these changepoints. This panel plots a posterior predictive check (Gelman et al. 2003) for overdispersion, which the model clearly fails. This plot shows a histogram of the standard deviations of data predicted by the posterior distribution of the parameters, along with the actual standard deviation of the data in red. Obviously the true standard deviation is considerably higher than what is predicted by the model. This is a clear indication that Poisson changepoint models have difficulty in situations where count data is overdispersed, such as with campaign contributions data.

3.4.2 The rise and fall of Herman Cain

Herman Cain's campaign for the 2012 Republican Presidential nomination provides an excellent demonstration of the validity of the above model. Cain was one of many candidates vying for the nomination and one of a few to reach the status of frontrunner, then quickly losing that status due in

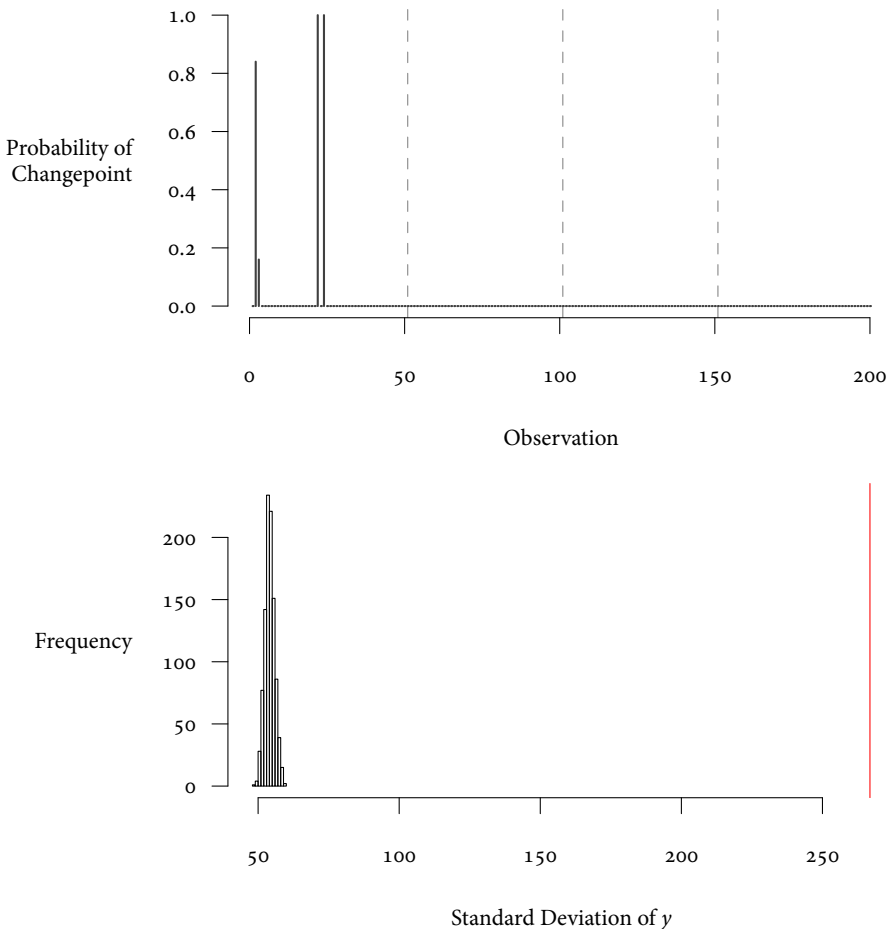


Figure 3.3: The results of a Poisson changepoint model with the simulated data. The top panel plots the posterior probability of a changepoint, with the true changepoints in dashed vertical lines. The bottom panel shows a posterior predictive check on the Poisson model, with a histogram of the standard deviations predicted by the posterior and the true standard deviation of the data in red.

part to allegations of sexual misconduct. The ups and downs of Cain's campaign provide a good target for the changepoint model.

To estimate this model, I use the above MCMC sampler with 100,000 iterations, thinned by 100, with a burn-in period of 5,000 iterations. Figure 3.4 presents the posterior probability of a changepoint in the top panel. In the bottom panel, I plot the raw number of contributors along with the posterior mean of

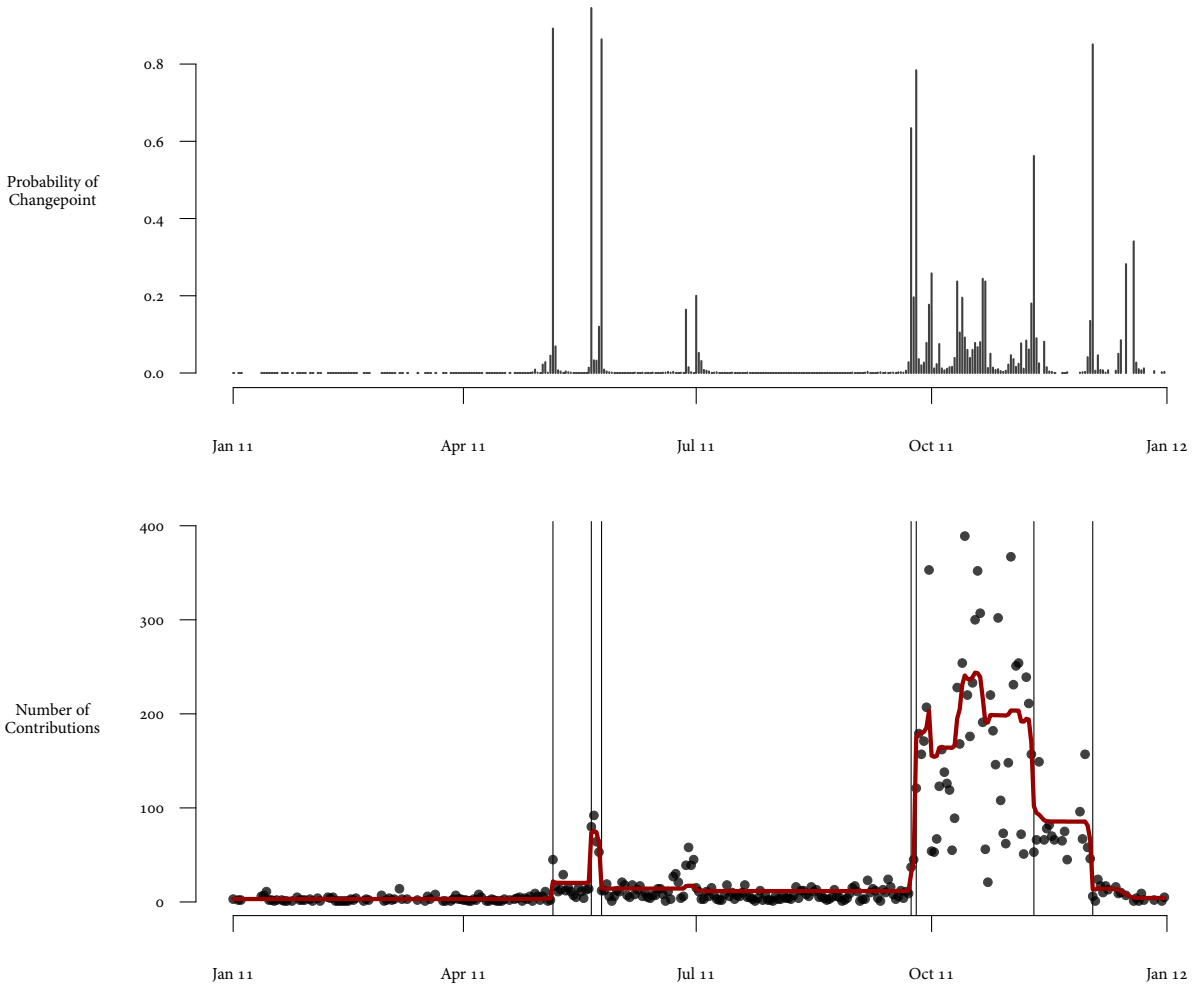


Figure 3.4: Contributions and changepoints for Herman Cain in the 2012 Republican Primary

λ_t, the mean of the negative binomial distribution for each observation in red. The vertical red lines correspond to dates that have greater than 0.5 posterior probability of being a changepoint. Table 3.1 lists each of these estimated changepoints and its corresponding campaign event in the campaign.

Although Cain officially announced his candidacy on May 21, 2011, he did participate in campaign activities before that time, including a Fox News debate on May 5th, where at least one Fox News focus

Estimated changepoint	Pr(Change)	Direction	Campaign Event
May 6, 2011	0.892	+	Fox News debate (May 5)
May 21–25, 2011	0.945	+	Announces candidacy (May 21)
September 23–25, 2011	0.784	+	Wins Florida 5 Straw Poll (Sept. 24)
November 11, 2011	0.562	−	Sexual misconduct allegations (Nov 7)
December 3, 2011	0.851	−	Suspends campaign (Dec. 3)

Table 3.1: Estimated Herman Cain changepoints and their substantive explanations.

group voted him the "winner." The model predicts a changepoint the day after this debate along with a short regime of high activity after he officially announces his candidacy. The model then estimates a long summer regime of June until late September when the model finds a series of changepoints following Cain's winning of the Florida 5 Straw Poll (Sutton and Holland 2011). This regime of increased contributions lasts a little over a month until November 10th, a little over a week after the first reports of Cain's sexual misconduct on October 30th (Martin et al. 2011) and a few days after the first women to go public with accusations against Cain on November 7th (Henderson 2011). This decidedly lower regime is ended by an estimated changepoint on the day that Cain suspends his campaign for the nomination.

The model correctly identify major shifts in the distribution of contributions to Herman Cain which correspond to actual prominent events in his campaign. It is important to note that the model makes no restrictions on the number of changepoints in the data. This is crucial in this example, because it is difficult to state the number of changepoint, even if one were to visually inspect the time series.

3.4.3 The senatorial surges

We can fruitfully apply this changepoint model to campaigns other than those at the national level. Furthermore, investigating local campaigns can give us insight into the relationship between local and national politics. To demonstrate this, I applied the gamechangers model to the fundraising for major-party nominees for Senate in the 2007-2008 election cycle.[3] One approach to modeling multiple

3. Due to small sample sizes, I dropped any candidate that had fewer than 200 contributors or fewer than 100 days of positive contributions. This left 59 out of a possible 68 candidates in 34 races. Note that there were two states, Mississippi and Wyoming, who had two Senate elections in 2008 due to special elections to replace vacated seats.

candidates at the same time would be to build a hierarchical version of the gamechangers model and run this larger model on all the candidates at the same time. This approach is slightly problematic for a changepoint problem. Namely, there is no reason, *a priori*, to think that the regimes of one campaign are necessarily the same fundamental type as regimes from another campaign. That is, it makes no sense to use the parameters from regime 2 in one race to help estimate the parameters in regime 2 in another race, since (a) regime 2 might be 1 day in one race and 100 days in the other or (b) regime 2 might not even occur in one of the races. To avoid this, we would have to either fix the number of changepoint across campaigns or implement a fairly complicated prior structure. Instead, I take the conceptually simpler approach and run the gamechangers model separately on each campaign. As in Section 3.4.2, I draw 100,000 MCMC iterations, thinned by 100, after an initial burn-in period of 5,000 draws.

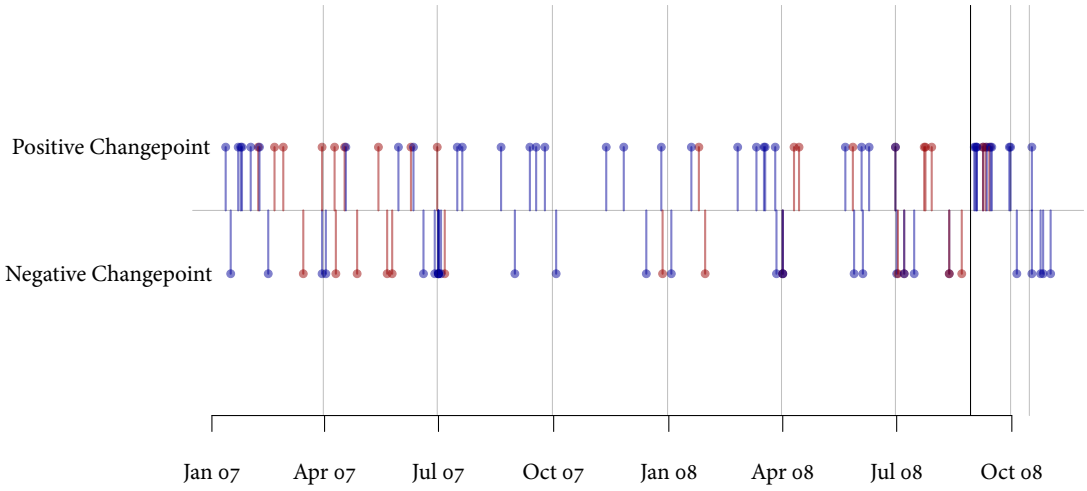


Figure 3.5: Changepoints for Senate races in 2008. The light grey lines are FEC filing deadlines. The black vertical line is the end of the Democratic National Convention.

The results from these models are presented in Figure 3.5, with the dates of Democratic changepoints in blue and Republican changepoints in red. In addition, the figure indicate the direction of the

changepoint depending on the sign of $\hat{\lambda}_t - \hat{\lambda}_{t-1}$, where t is the changepoint and $\hat{\lambda}_t$ is the mean of the posterior mean of the negative binomial at time t. The broad strokes of these results present an interesting picture. There is a flurry of activity early in the election cycle, then a relative calm in late 2007, then a steady pace in 2008. It is interesting to note that incumbent candidates dominate the "early money" gamechangers: 43 of the 48 changepoints in 2007 are for incumbent candidates (the changepoints in 2008 are almost exactly evenly divided between incumbents and non-incumbents).

In addition to locating the changepoints for each race, the game-changers model allows us to identify candidates who have certain type of changepoints. For instance, we may be interested in "surging" candidates: those whose fundraising takes off toward the end of the race. It is useful to identify these candidates, because they may give us insight as to how elites choose to contribute in close races.

State	Candidate	Changepoint	% vote	CQ (Spring)	CQ (Fall)
AK	Begich (D)	Sep. 29, 2008	50.66	Lean R	Leans D
CO	Udall (D)	Sep. 01, 2008	55.41	Tossup	Leans D
MN	Franken (D)	Sep. 03, 2008	50.01	Tossup	Tossup
NC	Hagan (D)	Sep. 11, 2008	54.37	Likely R	Tossup
NH	Shaheen (D)	Sep. 08, 2008	53.27	Tossup	Likely D
OR	Merkley (D)	Sep. 03, 2008	51.77	Lean R	Tossup

Table 3.2: Senate candidates who surged in 2008, as determined by the changepoint model. The "% vote" column is the their share of the two-party vote on election day. The CQ scores are the predictions made by *Congressional Quarterly* about the race in the Spring and the Fall.

To identify the surgers, I find all the campaign that had changepoints from September 1st, 2008 onward and that reached their maximum average contributions in the two months before election day. Table 3.2 shows the campaigns that meet this criteria in 2008, along with predictions from *Congressional Quarterly* and the final election outcome. Of these, five candidates are Democrats facing Republican incumbents, with only Rep. Mark Udall (CO) running for an open seat. All of the surging candidates had Spring predictions were either tossups or favoring the Republican. By October, the CQ rating had either remained the same or now favored the Democrat in each of the races. Furthermore,

each of these candidates ended up wining their race, albeit sometimes by small margins.

Interestingly, all of these candidates were identified by various Democratic fundraising groups as being targets for overturning Republican-held seats. During the month of September, former Vice-President Al Gore sent emails to members of the liberal group MoveOn.org to encourage them to donate to the campaigns of Hagan, Franken, and Udall (Davis 2008). Early in September, a group of prominent Hollywood women organized a group called "Voices for a Senate Majority" which sought to raise at least $100,000 for each of these candidates (Ressner 2008). The ability of candidates to raise funds is often thought of as critical and investigating why and how certain candidates are able to surge in such contribution toward the end of the race could bring valuable insights into the causes and consequences of campaign contributions more broadly. The game-changers models allows this kind of study by identifying these surging campaigns.

3.4.4 Game-changers and news coverage

Now that we have estimated changepoints for each Senate candidates for 2008, we may wish to understand what relationship these game-changers have with other aspects of the campaign. One way in which periods with changepoints differ from periods without changepoints is in the how the press covers the them. To demonstrate this, I collected data on the amount of coverage dedicated to each Senate race in each week using a political trade publication called *The Bulletin's Frontrunner*. This daily publications provides summaries of the national and local news coverage of each race. To measure the amount of coverage, I count the number of words in these summaries aggregated up the weekly level. This measure varies from zero words in some weeks to up to roughly 3,000 words toward the end of the campaign.

To get a sense for how game-changers relate to news coverage, I ran a logistic regression of the presence of a changepoint in a given race in a given week on the number of words written about that race in the *Frontrunner*. In addition, I included a linear time trend, the number of ads run by the

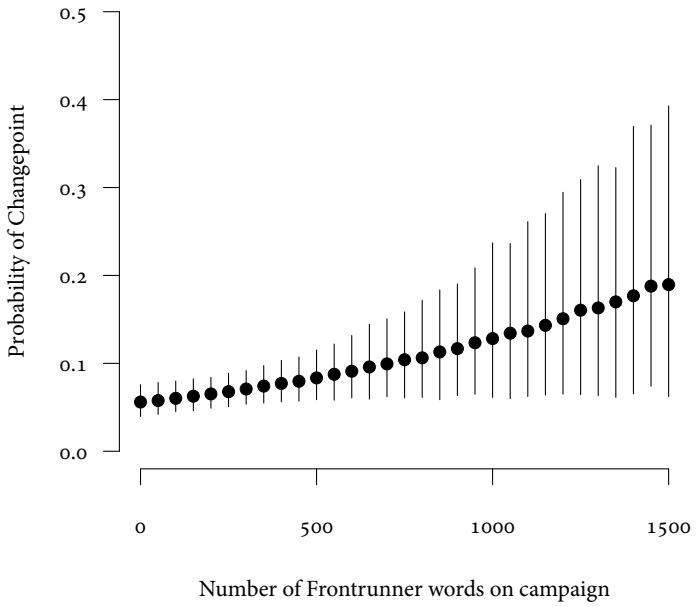


Figure 3.6: Probability of a changepoint as a function of the news coverage as measured by the *Frontrunner* word count.

candidates or the parties in that week, the Democratic percent of two-party poll results in that week, and, in some specifications, a race fixed-effect. Figure 3.6 shows how the probability of a changepoint changes with the *Frontrunner* word count.[4] In weeks with more news coverage, there is a greater chance of a changepoint and moving from 0 words to 1000 words roughly double those chances. This result seems to indicate that the dynamics of contributions might depend heavily on the attention paid to a campaign.

4. This was generated by the simulation-based marginal effects method of King, Tomz, and Wittenberg (2000), using `Zelig` (Imai, King, and Lau 2006).

3.5 Conclusion

Some campaigns take off and some campaigns fall flat. This essay presents a novel statistical model that estimates the number and timing of these changepoints in campaign contributions data. This model gives researchers the ability to detect significant events in campaigns and investigate the nature of these shifts in the broader political context. This represents the first attempt to measure a fairly tricky, yet common phenomenon: a campaign game-changer. With the game-changers model in hand, we can estimate changepoints for a whole host of campaigns and for a whole host contribution types—individuals versus PACs, men versus women, or in-state versus out-of-state. Further exploring the variation in structural breaks will help us better understand the nature of contributions as political participation.

Methodologically, the game-changer model pushes changepoint models forward by bringing together a few novel features. First, it naturally incorporates the overdispersion that is common in count data. Second, it leans on Bayesian nonparametrics in order to estimate the number of changepoint instead of having to know it *a priori*. This second contribution is especially important when, as in this case, marginal likelihoods are difficult to compute. One obvious way to extend this model is to build a multivariate version of the game-changers model. This model would estimate the changepoints for multiple time-series at the same time, allowing for in-model comparisons and complicated dependence structures. A potentially useful approach might be to combine the present model with the dynamic overdispersion model of Brandt and Sandler (2012).

The Dirichlet process prior approach that I take in this essay is more general than this specific negative binomial outcome model. Because it generalizes the Chib (1998) method for multiple changepoints, it also inherits the broad applicability of that method. Since the model parameters Θ are drawn conditional on the latent states and the Dirichlet process prior only affects the drawing of the latent state, it is straightforward to adapt this approach to changepoint model for continuous, binary, and ordered categorical variables such as those in Park (2011) or Spirling (2007).

A

Appendix to "Multiple Overimputation for Missing Data and Measurement Error"

Here we introduce a general MO model and a specific EM algorithm implementation to this end. We also show that it is equivalent to MI with observation-level priors as introduced by Honaker and King 2010. We also offer more general notation than that in the text.

A.1 General Framework

Consider a data set with independent and identically distributed random vectors $x_i = (x_{i1}, \ldots, x_{ip})$ with $i \in \{1, \ldots, N\}$. We are interested in the distribution of x_i, yet we only observe a distorted version of it, y_i. Let θ refer to the unknown parameters of the ideal data and γ refer to those of the error distribution. Thus, we have distributions $p(x_i|\theta)$ and $p(y_i|x_i, \gamma)$. As with MI, our goal is to produce copies of the ideal data, x_i, based on the observed data y_i.

We define $e_i = (e_{i1}, \ldots, e_{ip})$ to be a vector of error indicators. The typical element e_{ij} takes a value of 1 to indicate that variable j on observation i is measured with error so that we observe a proxy, $y_{ij} = w_{ij}$ instead of x_{ij}. Similarly, we define m_i to be a vector of missingness indicators. When m_{ij} takes the value 1, then y_{ij} is missing. If both $m_{ij} = 0$ and $e_{ij} = 0$, then the observation is perfectly measured and $y_{ij} = x_{ij}$. Let m_i and e_i have a joint distribution $p(m_i, e_i|y_i, x_i, \varphi)$, whose parameters φ are distinct from θ and γ.

With these definitions in hand, we can decompose each observation into various subsets. Let x_i^{obs} be all the perfectly measured values, so that $x_i^{\text{obs}} = \{x_{ij};\ e_{ij} = m_{ij} = 0\}$. We also have x_i^{mis}, which are the variables that are missing in observation i: $x_i^{\text{mis}} = \{x_{ij};\ m_{ij} = 1\}$. Finally, we must define those variables that are measured with error. Let x_i^{err} be the unobserved, latent variables and w_i be their observed proxies: $w_i = \{w_{ij};\ e_{ij} = 1\}$, $x_i^{\text{err}} = \{x_{ij};\ e_{ij} = 1\}$. Thus, the observed data for any unit will be $y_i = (x_i^{\text{obs}}, w_i)$ and the ideal data would be $x_i = (x_i^{\text{obs}}, x_i^{\text{err}}, x_i^{\text{mis}})$. Note that while the dimensions of x_i and y_i are fixed, the dimensions of both w_i and x_i^{obs} can change from unit to unit.

We can write the observed-data probability density function for unit i as

$$p(y_i, m_i, e_i|\theta, \gamma, \varphi) = \int\int p(x_i|\theta)p(w_i|x_i, \gamma)p(m_i, e_i|y_i, x_i, \varphi)dx_i^{\text{err}}dx_i^{\text{mis}}. \tag{A.1}$$

We make the assumption that the data is mismeasured at random (MMAR), which states that the

mismeasurement and missingness processes do not depend on the unobserved data.[1] Formally, we state MMAR as $p(m_i, e_i|y_i, x_i, \varphi) = p(m_i, e_i|y_i, \varphi)$. With this assumption in hand, we can rewrite (A.1) as $p(y_i, m_i, e_i|\theta, \gamma, \varphi) = p(m_i, e_i|y_i, \varphi)p(y_i|\theta, \gamma)$, and since we are primarily interested in inferences on θ, the first term becomes part of the proportionality constant and we are left with the observed-data distribution

$$p(y_i|\theta, \gamma) = \int \int p(x_i|\theta)p(w_i|x_i, \gamma)dx_i^{\text{err}}dx_i^{\text{mis}}. \tag{A.2}$$

Taking a Bayesian point of view, we can combine this with a prior on (θ, γ) giving us a posterior, $p(\theta, \gamma|y_i)$.

Analyzing the ideal data x_i would be much easier than y_i since the mismeasured and missing data contribute to likelihood in complicated ways. Thus, MO seeks to form a series of complete, ideal data sets: $x_{i(1)}, x_{i(2)}, \ldots, x_{i(m)}$. Each of these overimputed data sets is of the form $x_{i(k)} = (x_i^{\text{obs}}, x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}})$, so that the perfectly measured data is constant across the overimputations. We refer to this as overimputation because we replace observed data w_i with draws from an imputation model for x_i^{err}. To form these overimputations, we take draws from the posterior predictive distribution of the unobserved data:

$$(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}}) \sim p(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}}|y_i) = \int p(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}}|y_i, \theta, \gamma)p(\theta, \gamma|y_i)d\theta d\gamma. \tag{A.3}$$

Once we have these m overimputations, we can simply run m separate analyses on each data set and combine them using straightforward rules. Consider some quantity of interest, Q. Let $q_1, \ldots, q_m$ denote the separate estimates of Q which come from applying the same analysis model to each of the overimputed data sets. The overall point estimate $\bar{q}$ of Q is simply the average $\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j$. As shown by Rubin, 1978, the variance of the multiple overimputation point estimate is the average of the estimated variances from within each completed data set, plus the sample variance in the point

1. This is an augmented version of the *missing at random* (MAR) assumption (Rubin 1976). MMAR would be violated if the presence of measurement error depended on the value of the latent variable itself. Since we have mismeasured proxies included in y_i, the dependence would have to be after controlling for the proxies. The most likely violation of this assumption would be if follow-up data were collected on certain observations that were different on some unmeasured covariate.

estimates across the data sets (multiplied by a factor that corrects for bias because $m < \infty$): $\bar{s}^2 = \frac{1}{m}\sum_{j=1}^{m} s_j^2 + S_q^2(1 + 1/m)$, where s_j is the standard error of the estimate of q_j from the analysis of data set j and $S_q^2 = \sum_{j=1}^{m}(q_j - \bar{q})/(m-1)$.[2]

A.2 A Modified-EM Approach to Multiple Overimputation

The last formulation of (A.3) hints at one way to draw multiple imputations: (1) draw $(\theta_{(i)}, \gamma_{(i)})$ from its posterior $p(\theta, \gamma|y_i)$, then (2) draw $(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}})$ from $p(x_{i(k)}^{\text{err}}, x_{i(k)}^{\text{mis}}|y_i, \theta_{(i)}, \gamma_{(i)})$. Usually these procedures are implemented with either data augmentation (that is, Gibbs sampling) or the expectation-maximization (EM) algorithm combined with an additional sampling step. We focus here on how our method works in the EM algorithm, since these two approaches are closely linked and often lead to similar inferences (Schafer 1997; King et al. 2001; Honaker and King 2010). EM consists of two steps: the expectation step, when we use the current guess of the parameters to fill in the missing data, and the maximization step, when we use the observed data and our current guess of the missing data to estimate the complete-data parameters. These two steps iterate until the parameters estimates converge.

If the mismeasured cells were in fact missing, we could easily apply a typical EM algorithm for missing data. In this case, though, the observed proxies, w_i, give us observation-level information about x_i^{err}. The EM algorithm usually incorporates prior beliefs about the parameters in the M-step, which is convenient when our prior beliefs are on the parameters of the data (μ, Σ). Here our information is about the location of a missing value, not about the parameters themselves.

We therefore include this information in the expectation- or E-step of the EM algorithm. This step calculates the expected value of the complete-data sufficient statistics over the full conditional distribution of the missing data. That is, it finds $E(T(x_i)|y_i, \theta^{(t)}, \gamma)$, where $\theta^{(t)}$ is the current guess of the complete-data parameters. In our model, we adjust the E-step to incorporate the measurement error

2. A second procedure for combining estimates is useful when simulating quantities of interest, as in King, Tomz, and Wittenberg (2000) and Imai, King, and Lau (2008). To draw m simulations of the quantity of interest, we merely draw $1/m$ of the needed simulations from each of the overimputed data sets.

distribution as implied by the observed-data likelihood, (A.2). Using this likelihood, the modified E-step calculates

$$E(T(x_i)|y_i, \theta^{(t)}, \gamma) = \int\int T(x_i) \underbrace{p(x_i^{\text{err}}, x_i^{\text{mis}}|x_i^{\text{obs}}, \theta^{(t)})}_{\text{imputation}} \underbrace{p(w_i|x_i, \gamma)}_{\text{mismeasurement}} dx_i^{\text{err}} dx_i^{\text{mis}}, \tag{A.4}$$

where in typical missing data applications of EM, the mismeasurement term would be absent. The imputation part of the expectation draws information from a regression of the missing data on the observed data, while the mismeasurement part draws information from the proxy.[3] Thus, both sources of information help estimate the true sufficient statistics of the latent, ideal data. The M-step proceeds as usual, finding the parameters that were most likely to have give rise to the estimated sufficient statistics. Note that we could incorporate this alteration to the full conditional posterior into an MCMC approach, though instead of averaging across the distribution, a Gibbs sampler would take a draw from it.

A.3 A Multiple Overimputation Model for Normal Data

In the above description of the model, we have left the distributions unspecified. To implement the model, we must provide additional information. We assume that the complete, ideal data (x_i) is multivariate normal with mean μ and covariance Σ, so that $\theta = (\mu, \Sigma)$. This implies that any conditional distribution of the ideal is also normal.

The above measurement error distribution is in its most general form, a function of the entire ideal data vector (x_i) and some parameters, γ. As noted by Stefanski (2000), all approaches to correcting measurement error must include additional information about this distribution. We assume that $w_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(x_{ij}, \lambda_{ij}^2)$ for each proxy $w_{ij} \in w_i$ and each unit i, where the measurement error variance λ_{ij}^2 is known or estimable using techniques from Section 2.3. Our assumption corresponds to that of classical measurement error, yet our modified EM algorithm can handle more general cases than this. If the

3. Note that we treat γ as fixed since, in our implementation, it is known or estimable. One could extend these methods to simultaneously estimate γ, though this would require additional information.

measurement error is known to be biased or dependent upon another variable, we can simply adjust the cell-level means above and proceed as usual. Essentially, one must have knowledge of *how* the variable was mismeasured. The simulation results in Section 2.4.2 further indicate that MO is robust to these assumptions in certain situations.

With the measurement error model above, the normality of the data makes the calculation of the sufficient statistics straightforward. To ease exposition, we assume that there are no missing values, so that $x_i^{\text{mis}} = \emptyset$. With only measurement error, the E-step becomes

$$E(T(x_i)|y_i, \theta^{(t)}) = \int T(x_i)\; p(x_i^{\text{err}}|x_i^{\text{obs}}, \theta^{(t)}) \prod_{w_{ij} \in w_i} p(w_{ij}|x_{ij}, \lambda_{ij}^2) dx_i^{\text{err}}, \tag{A.5}$$

where $T(x_i)$ is the set of sufficient statistics for the multivariate normal. In a slight abuse of notation, we can gather the independent measurement error distributions, w_i, into a multivariate normal with mean x_i^{err} and covariance matrix $\Lambda_i = \lambda_i^2 I$, where $\lambda_i^2 = \{\lambda_{ij}^2;\; e_{ij} = 1\}$ and I is the identity matrix with dimension equal to $\sum_j e_{ij}$.

In order to calculate the expectation in (A.5), we must know the full conditional distribution, which is $p(x_i^{\text{err}}|y_i, \theta, \lambda_i^2) \propto p(x_i^{\text{err}}|x_i^{\text{obs}}, \theta) p(w_i|x_i^{\text{err}}, \lambda_i^2)$. Note that each of the distributions is (possibly multivariate) normal, with $x_i^{\text{err}}|x_i^{\text{obs}}, \theta \sim \mathcal{N}(\mu_{\text{e|o}}, \Sigma_{\text{e|o}})$ and $w_i|x_i^{\text{err}}, \lambda_i^2 \sim \mathcal{N}(x_i^{\text{err}}, \Lambda_i)$, where $(\mu_{\text{e|o}}, \Sigma_{\text{e|o}})$ are deterministic functions of θ and x_i^{obs}. This distribution amounts to the regression of x_i^{err} on x_i^{obs}. If the values were simply missing, rather than measured with error, then the E-step would simply take the expectations with respect to this conditional expectation. With measurement error, we must combine these two sources of information. Using standard results on the normal distribution, we can write the full conditional as

$$(x_i^{\text{err}}|y_i, \theta^{(t)}, \lambda_i^2) \sim \mathcal{N}(\mu^*, \Sigma^*), \quad \Sigma^* = (\Lambda_i^{-1} + \Sigma_{\text{e|o}}^{-1})^{-1}, \quad \mu^* = \Sigma^*(\Lambda_i^{-1} w_i + \Sigma_{\text{e|o}}^{-1} \mu_{\text{e|o}}). \tag{A.6}$$

We simply change our E-step to calculate this expectation for each cell measured with error and proceed

with the M-step as usual.[4] Note that while we assume that the measurement errors on different variables are independent, one could incorporate dependence into Λ_i. The result in (A.6) is identical to the results in the appendix of Honaker and King 2010, when we set a prior distribution for x_i^{err} that is normal with mean w_i and variance Λ_i. See their paper for additional implementation details.

4. If there are missing values in unit i, we need to alter the definitions of Λ_i^{-1} and w_i to be 0 for the entries corresponding to the missing variables.

Bibliography

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: estimating the effect of california's tobacco control program. *Journal of the American Statistical Association* 105 (490): 493–505. http://pubs.amstat.org/doi/abs/10.1198/jasa.2009.ap08746. (Cit. on p. 9).

Achen, Christopher. 1986. *Statistical analysis of quasi-experiments.* Berkeley: University of California Press. http://books.google.com/books?id=Qbk8XgU57aQC. (Cit. on p. 16).

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* 91:444–455. http://www.jstor.org/stable/2291629. (Cit. on p. 16).

Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder. 2008. The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. *American Political Science Review* 102 (02). (Cit. on pp. 75, 76).

Bartels, Larry M. 1985. Expectations and preferences in presidential nominating campaigns. *American Political Science Review*:804–815. (Cit. on p. 82).

Berger, James. 1994. An overview of robust bayesian analysis (with discussion). *Test* 3:5–124. (Cit. on p. 43).

Bityukov, SI, VV Smirnova, NV Krasnikov, and VA Taperechkina. 2006. Statistically dual distributions in statistical inference. In *Statistical problems in particle physics, astrophysics and cosmology: proceedings of phystat05, oxford, uk, 12-15 september 2005*, 102–105. Http://arxiv.org/abs/math/0411462v2. (Cit. on p. 42).

Black, Dan A., Mark C. Berger, and Frank A. Scott. 2000. Bounding parameter estimates with nonclassical measurement error. *Journal of the American Statistical Association* 95 (451): 739–748. http://www.jstor.org/stable/2669454. (Cit. on p. 49).

Brandt, Patrick T., and Todd Sandler. 2012. A bayesian poisson vector autoregression model. *Political Analysis*. http://pan.oxfordjournals.org/content/early/2012/03/15/pan.mps001.abstract. (Cit. on p. 104).

Brownstone, David, and Robert G. Valletta. 1996. Modeling Earnings Measurement Error: A Multiple Imputation Approach. *Review of Economics and Statistics* 78 (4): 705–717. (Cit. on p. 38).

Carroll, Raymond J., and Leonard A. Stefanski. 1990. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* 85 (411): 652–663. http://www.jstor.org/stable/2290000. (Cit. on p. 48).

Carroll, R.J., D. Ruppert, and L.A. Stefanski. 1995. *Measurement error in nonlinear models.* Vol. 63. Chapman & Hall/CRC. (Cit. on p. 48).

Chib, Siddhartha. 1995. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association* 90 (432): pp. 1313–1321. http://www.jstor.org/stable/2291521. (Cit. on p. 88).

———. 1998. Estimation and comparison of multiple change-point models. *Journal of Econometrics* 86 (2): 221–241. (Cit. on pp. 82, 85, 86, 88, 90, 92, 93, 104).

Chib, Siddhartha, and Edward Greenberg. 1996. Markov chain monte carlo simulation methods in econometrics. *Econometric Theory* 12 (3): pp. 409–431. http://www.jstor.org/stable/3532527. (Cit. on p. 86).

Chong, Terence Tai-Leung, and Stanley Iat-Meng Ko. 2011. Dirichlet process multiple change-point model. Presented at the Econometric Society Australasian Meeting 2011, Adelaide, Australia. (Cit. on p. 88).

Christenson, Dino P, and Corwin D Smidt. 2011. Riding the Waves of Money: Contribution Dynamics in the 2008 Presidential Nomination Campaign. *Journal of Political Marketing* 10, nos. 1-2 (Feb.): 4–26. (Cit. on p. 84).

Cole, Stephen R., Haitao Chu, and Sander Greenland. 2006. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* 35 (4): 1074–81. (Cit. on pp. 38, 48).

Cole, Stephen R., and Constantine E. Frangakis. 2009. The consistency statement in causal inference: a definition or an assumption? *Epidemiology* 20 (1): 3–5. (Cit. on p. 11).

Cole, Stephen R., and Miguel A Hernán. 2008. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiolology* 168 (6): 656–64. (Cit. on p. 27).

Cook, J., and L. Stefanski. 1994. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Asociation* 89:1314–1328. (Cit. on p. 49).

Davis, Susan. 2008. Gore, moveon team up to aid senate candidates. *Wall Street Journal* (Sept. 7). http://blogs.wsj.com/washwire/2008/09/18/gore-moveon-team-up-to-aid-senate-candidates/. (Cit. on p. 102).

Escobar, Michael D., and Mike West. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90 (430): pp. 577–588. http://www.jstor.org/stable/2291069. (Cit. on p. 89).

Federal Election Commission. 2011. *Campaign guide for congressional candidates and committees.* (Cit. on p. 83).

Ferguson, Thomas S. 1973. A bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1 (2): pp. 209–230. http://www.jstor.org/stable/2958008. (Cit. on p. 89).

Frangakis, Constantine E., and Donald B. Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58:21–29. http://www.jstor.org/stable/3068286. (Cit. on p. 9).

Freedman, Laurence S, Douglas Midthune, Raymond J Carroll, and Victor Kipnis. 2008. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine* 27 (25): 5195–216. (Cit. on p. 38).

Frühwirth-Schnatter, Sylvia, Rudolf Frühwirth, Leonhard Held, and Håvard Rue. 2009. Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing* 19 (4): 479–492. (Cit. on pp. 87, 91, 93, 94).

Fuller, Wayne A. 1987. *Measurement error models.* Wiley New York. (Cit. on pp. 48, 78).

Gelman, Andrew, J.B. Carlin, H.S. Stern, and D.B. Rubin. 2003. *Bayesian data analysis, second edition.* Chapman & Hall. (Cit. on p. 96).

Geweke, John, and Yu Jiang. 2011. Inference and prediction in a multiple-structural-break model. *Journal of Econometrics* 163 (2): 172–185. (Cit. on p. 88).

Ghosh-Dastidar, Bonnie, and Joseph L Schafer. 2003. Multiple Edit/Multiple Imputation for Multivariate Continuous Data. *Journal of the American Statistical Association* 98 (464): 807–817. (Cit. on p. 38).

Giordani, Paolo, and Robert Kohn. 2008. Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models. *Journal of Business & Economic Statistics* 26, no. 1 (Jan.): 66–77. (Cit. on p. 88).

Glynn, Adam N. 2011. The product and difference fallacies for indirect effects. *American Journal of Political Science:*no–no. http://dx.doi.org/10.1111/j.1540-5907.2011.00543.x. (Cit. on p. 9).

Glynn, Adam N., and Kevin M. Quinn. 2010. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis* 18 (1): 36–56. http://pan.oxfordjournals.org/content/18/1/36.abstract. (Cit. on p. 8).

Goldstein, Kenneth, and Joel Rivlin. 2007. *Congressional and gubernatorial advertising, 2003-2004.* Combined File [dataset]. Final release. Madison, WI. http://wiscadproject.wisc.edu/. (Cit. on p. 21).

Grimmer, Justin. 2011. An introduction to bayesian inference via variational approximations. *Political Analysis* 19 (1): 32–47. http://pan.oxfordjournals.org/content/19/1/32.abstract. (Cit. on p. 89).

Guolo, Annamaria. 2008. Robust techniques for measurement error correction: a review. *Statistical Methods in Medical Research* 17 (6): 555–80. (Cit. on p. 36).

Heckman, James. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables, and simple estimator for such models. *Annals of Economic and Social Measurement* 5:475–492. http://www.nber.org/chapters/c10491. (Cit. on p. 16).

Henderson, Nia-Malika. 2011. Sharon bialek accuses herman cain of sexual harassment as she sought help getting a job. *Washington Post* (Nov. 7). http://www.washingtonpost.com/politics/sharon-bialek-accuses-cain-speaks-out/2011/11/07/gIQADYPlvM_story.html. (Cit. on p. 99).

Hernán, Miguel A, Emilie Lanoy, Dominique Costagliola, and James M. Robins. 2006. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology Toxicology* 98 (3): 237–42. (Cit. on p. 33).

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2006. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 15 (3): 199. (Cit. on pp. 5, 26).

Honaker, James, and Gary King. 2010. What to do about missing values in time series cross-section data. HTTP://GKING.HARVARD.EDU/FILES/ABS/PR-ABS.SHTML, *American Journal of Political Science* 54, no. 2 (Apr.): 561–581. (Cit. on pp. 38, 45, 70, 105, 108, 111).

Honaker, James, Gary King, and Matthew Blackwell. 2010. Amelia ii: a program for missing data. HTTP://GKING.HARVARD.EDU/AMELIA. (Cit. on p. 37).

Honaker, James, and Eric Plutzer. 2011. Small area estimation with multiple overimputation. Paper presented at the Midwest Political Science Association, Chicago. (Cit. on p. 72).

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe [in English]. *Journal of the American Statistical Association* 47 (260): pp. 663–685. http://www.jstor.org/stable/2280784. (Cit. on p. 8).

Huckfeldt, Robert, Eric Plutzer, and John Sprague. 1993. Alternative contexts of political behavior: churches, neighborhoods, and individuals. *Journal of Politics* 55, no. 2 (May): 365–381. (Cit. on pp. 70, 72).

Imai, Kosuke, Gary King, and Olivia Lau. 2006. Zelig: everyone's statistical software. HTTP://GKING.HARVARD.EDU/ZELIG. (Cit. on p. 103).

Imai, Kosuke, Gary King, and Olivia Lau. 2008. Toward a common framework for statistical analysis and development. HTTP://GKING.HARVARD.EDU/FILES/ABS/Z-ABS.SHTML, *Journal of Computational Graphics and Statistics* 17 (4): 1–22. (Cit. on p. 108).

Imai, Kosuke, and Teppei Yamamoto. 2010. Causal inference with differential measurement error: nonparametric identification and sensitivity analysis. *American Journal of Political Science* 54, no. 2 (Apr.): 543–560. (Cit. on pp. 41, 79).

Ishwaran, Hemant, and Lancelot F. James. 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 96 (453): 161–173. (Cit. on p. 90).

Katz, Jonathan N., and Gabriel Katz. 2010. Correcting for Survey Misreports Using Auxiliary Information with an Application to Estimating Turnout. *American Journal of Political Science* 54 (3): 815–835. (Cit. on p. 79).

King, Gary. 1989. Variance specification in event count models: from restrictive assumptions to a generalized estimator. HTTP://GKING.HARVARD.EDU/FILES/ABS/VARSPECEC-ABS.SHTML, *American Journal of Political Science* 33, no. 3 (Aug.): 762–784. (Cit. on pp. 82, 87).

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. Making the most of statistical analyses: improving interpretation and presentation. HTTP://GKING.HARVARD.EDU/FILES/ABS/MAKING-ABS.SHTML, *American Journal of Political Science* 44, no. 2 (Apr.): 341–355. (Cit. on pp. 103, 108).

King, Gary, and Langche Zeng. 2002. Estimating risk and rate levels, ratios, and differences in case-control studies. HTTP://GKING.HARVARD.EDU/FILES/ABS/1S-ABS.SHTML, *Statistics in Medicine* 21:1409–1427. (Cit. on p. 43).

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. HTTP://GKING.HARVARD.EDU/FILES/ABS/EVIL-ABS.SHTML, *American Political Science Review* 95, no. 1 (Mar.): 49–69. (Cit. on pp. 36, 39, 78, 108).

Klepper, Steven, and Edward E. Leamer. 1984. Consistent sets of estimates for regressions with errors in all variables. *Econometrica* 52 (1): 163–184. http://www.jstor.org/stable/1911466. (Cit. on p. 49).

Koop, Gary, and Simon M. Potter. 2007. Estimation and forecasting in models with multiple breaks. *Review of Economic Studies* 74 (3): 763–789. (Cit. on p. 88).

———. 2009. Prior Elicitation in Multiple Change-point Models. *International Economic Review* 50 (3): 751–772. (Cit. on p. 88).

Lau, Richard R., and Gerald M. Pomper. 2002. Effectiveness of Negative Campaigning in U.S. Senate Elections. *American Journal of Political Science* 46 (1): 47–66. http://www.jstor.org/stable/3088414. (Cit. on p. 20).

Lau, Richard R., and Gerald M. Pomper. 2004. *Negative Campaigning: An Analysis of U.S. Senate Elections.* Campaigning American Style. Lanham, MD: Rowman & Littlefield Publishers, Inc. http://books.google.com/books?id=56SobRCutpEC. (Cit. on pp. 20, 21, 26).

Lau, Richard R., Lee Sigelman, and Ivy Brown Rovner. 2007. The effects of negative political campaigns: a meta-analytic reassessment. *The Journal of Politics* 69 (04): 1176–1209. (Cit. on pp. 3, 20, 26).

Leamer, Edward. 1978. *Specification searches.* New York: Wiley. (Cit. on pp. 49, 51).

Lee, Sik-Yum. 2007. *Structural equation modeling: a bayesian approach.* Vol. 680. John Wiley & Sons Inc. (Cit. on p. 49).

Lumley, Thomas. 2004. Analysis of Complex Survey Samples. R package verson 2.2, *Journal of Statistical Software* 9 (1): 1–19. http://www.jstatsoft.org/v09/i08. (Cit. on p. 19).

———. 2010. survey*: analysis of complex survey samples.* R package version 3.23-2. http://faculty.washington.edu/tlumley/survey/. (Cit. on p. 25).

Martin, Jonathan, Maggie Haberman, Anna Palmer, and Kenneth P. Vogel. 2011. Herman cain accused by two women of inappropriate behavior. *Politico* (Oct. 31). http://www.politico.com/news/stories/1011/67194.html. (Cit. on p. 99).

Meng, Xiao-Li. 1994. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 9 (4): 538–573. (Cit. on p. 46).

Mutz, Diana C. 1995. Effects of horse-race coverage on campaign coffers: strategic contributing in presidential primaries. *The Journal of Politics* 57 (04): 1015–1042. http://dx.doi.org/10.1017/S0022381600050611. (Cit. on p. 82).

Neal, Radford M. 2000. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 9 (2): pp. 249–265. http://www.jstor.org/stable/1390653. (Cit. on pp. 89, 90).

Park, Jong Hee. 2010. Structural change in u.s. presidents' use of force. *American Journal of Political Science* 54 (3): 766–782. http://dx.doi.org/10.1111/j.1540-5907.2010.00459.x. (Cit. on pp. 82, 86, 88, 96).

———. 2011. Changepoint analysis of binary and ordinal probit models: an application to bank rate policy under the interwar gold standard. *Political Analysis* 19 (2): 188–204. http://pan.oxfordjournals.org/content/19/2/188.abstract. (Cit. on pp. 86, 88, 104).

Pearl, Judea. 2010. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology* 21 (6): 872–5. (Cit. on p. 11).

Pierson, Paul. 2000. Not Just What, but When: Timing and Sequence in Political Processes. *Studies in American Political Development* 14 (1): 72. (Cit. on p. 4).

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. http://www.R-project.org. (Cit. on p. 18).

Ressner, Jeffrey. 2008. Smile! you're in poliwood. *Politico* (Sept. 4). http://www.politico.com/news/stories/0908/13097.html. (Cit. on p. 102).

Robins, James M. 1986. A new approach to causal inference in mortality studies with sustained exposure periods-Application to control of the healthy worker survivor effect. *Mathematical Modelling* 7 (9-12): 1393–1512. http://biosun1.harvard.edu/~robins/new-approach.pdf. (Cit. on p. 9).

———. 1997. Causal Inference from Complex Longitudinal Data. In *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane, 69–117. Vol. 120. Lecture Notes in Statistics. New York: Springer-Verlag. http://biosun1.harvard.edu/~robins/cicld-ucla.pdf. (Cit. on pp. 9, 10, 15, 32, 33).

———. 1999. Association, Causation, and Marginal Structural Models. *Synthese* 121 (1/2): 151–179. {http://www.jstor.org/stable/20118224}. (Cit. on pp. 8, 29, 30).

———. 2000. Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, ed. M. Elizabeth Halloran and Donald Berry, 95–134. Vol. 116. The IMA Volumes in Mathematics and its Applications. New York: Springer-Verlag. http://biosun1.harvard.edu/~robins/msm-cie-fnl.pdf. (Cit. on pp. 8, 19).

Robins, James M., Miguel A Hernán, and Babette Brumback. 2000. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 11 (5): 550–560. http://www.jstor.org/stable/3703997. (Cit. on pp. 2, 13, 16, 19).

Rubin, Donald B. 1976. Inference and missing data. *Biometrika* 63:581–592. (Cit. on pp. 40, 107).

———. 1978. Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics* 6 (1): 34–58. http://www.jstor.org/stable/2958688. (Cit. on pp. 9, 10, 107).

———. 1987. *Multiple imputation for nonresponse in surveys*. New York: John Wiley. (Cit. on p. 36).

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. London: Chapman & Hall. (Cit. on pp. 38–40, 48, 108).

Schafer, Joseph L., and Maren K. Olsen. 1998. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research* 33 (4): 545–571. (Cit. on p. 48).

Spirling, Arthur. 2007. Bayesian approaches for limited dependent variable change point problems. *Political Analysis* 15 (4): 387–405. http://pan.oxfordjournals.org/content/15/4/387.abstract. (Cit. on pp. 82, 86, 104).

Spirling, Arthur, and Kevin Quinn. 2010. Identifying intraparty voting blocs in the u.k. house of commons. *Journal of the American Statistical Association* 105 (490): 447–457. http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.ap07115. (Cit. on p. 89).

Stefanski, L. A. 2000. Measurement Error Models. *Journal of the American Statistical Association* 95 (452): 1353–1358. (Cit. on pp. 36, 50, 109).

Sutton, Jane, and Steve Holland. 2011. Cain upsets perry in florida republican straw poll. *Reuters* (Sept. 24). http://www.reuters.com/article/2011/09/25/us-usa-campaign-winner-idUSTRE78N2RE20110925. (Cit. on p. 99).

VanderWeele, Tyler J. 2009. Concerning the consistency assumption in causal inference. *Epidemiology* 20 (6): 880–3. (Cit. on p. 11).

Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and equality: civic volunteerism in american politics.* Cambridge, MA: Harvard University Press. (Cit. on p. 82).

Wang, Naisyin, and James M. Robins. 1998. Large-sample theory for parametric multiple imputation procedures. *Biometrika* 85:935–948. (Cit. on p. 38).

White, Ian R. 2006. Commentary: dealing with measurement error: multiple imputation or regression calibration? *International Journal of Epidemiology* 35 (4): 1081–2. (Cit. on p. 48).

Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73 (1): 3–36. (Cit. on p. 22).

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX.
A template that can be used to format a PhD thesis with this look and feel is freely available online at https://github.com/suchow/.